UNIVERSITY OF
CAMBRIDGE

Development of a Multiscale Coarse-Grained Chromatin Model

Stephen Edward Farr

Department of Physics, University of Cambridge
Jesus College

April 2021

This thesis is submitted for the degree of
Doctor of Philosophy

Supervisor: Dr Rosana Collepardo-Guevara

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Abstract

An important challenge in understanding gene behavior is deciphering how the genome is organized in space and how this organization influences its function. Existing experimental and computational methods lack the ability to provide close up views of the structure of biomolecules inside nano-scale chromatin. In this thesis, we develop a multiscale coarse-grained chromatin model, which integrates all-atom representations of proteins, DNA, and nucleosomes; a chemically specific coarse-grained model of kb scale chromatin; and a minimal model of sub-Mb scale chromatin. A key feature of this model is its capacity to link the molecular details of nucleosomes to the collective behavior of mesoscale (up to sub-Mb scale) chromatin.

Our chemically-specific model describes DNA at base-pair resolution and proteins at amino-acid level resolution. We have used this model to investigate how sub-nucleosome level physicochemical and structural properties, such as the spontaneous thermal breathing and sliding motion of DNA, affect larger scale chromatin self-assembly. Nucleosome breathing refers to the observation that nucleosomes, rather than being static particles, exhibit spontaneous structural fluctuations where the DNA binds and unbinds dynamically. We find that such plasticity of nucleosomes destabilizes the highly regular zig-zag fiber chromatin folding configurations, and promotes instead an irregular and dynamical organization of nucleosomes termed 'liquid-like'. Our model can also be used to investigate the effects of DNA sequence, salt conditions, and binding of additional proteins on the behavior of chromatin.

Our minimal model describes nucleosomes with just a few particles, while still explicitly representing the DNA. We have used our minimal model to investigate the phase behavior of systems of multiple interacting chromatin fibers. We find that chromatin undergoes salt-mediated liquid-liquid phase separation, and that nucleosome plasticity plays an important role in increasing the range of stability of the coexistence region. Additionally, the model is able to investigate the size scaling properties of chromatin fibers and the effect of nucleosome repeat length on chromatin compaction and inter-chromatin interactions.

Together, our multiscale methodology provides a useful technique to extrapolate atomistic properties of nucleosomes to the modulation of large-scale chromatin organization.

Acknowledgments

Thanks to my supervisor Rosana Collepardo-Guevara for her support, guidance, and enthusiasm. It has been a pleasure to work alongside such a talented scientist. Thanks to all members of the Collepardo group for the support and friendship, with a special thanks to: Adiran Garaizar for his assistance with numerous aspects of computer simulation, Jorge Rene-Espinosa for interesting discussions, Jerelle Joseph for her advice on scientific writing, and Esmae Woods for checking my writing, methods, mathematics, and code.

Thanks to the EPSRC centre for Doctoral Training in Computational Methods for Materials Science for funding. Thanks to the Maxwell centre coffee machine.

Thanks to my friends and family for their support over the years, particularly for keeping me relatively sane during the final year of pandemic life. Special thanks to the original physics team: Joe, Milo, Sam, and Craig.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

To fit inside an eukaryotic cell, DNA undergoes phenomenal levels of packaging and compaction in a cell-cycle dependent manner. At interphase, the DNA is relatively disordered; i.e., under a microscope it appears as patchy regions of high and low density which are thought to correspond to inactive non-coding (heterochromatin) and active coding (euchromatin) regions of DNA respectively [1]. During cell division, the DNA becomes highly packaged and ordered, forming the familiar X-shape chromosomes [2]. To illustrate this level of packaging, we note that there are approximately 3 billion DNA base pairs (bp) in the human genome, or 6 billion during cell division, and that the separation among two consecutive base pairs is 3.4 Å. This results in a total length of approximately 2 m for the human genome. Given that the diameter of a cell nucleus is approximately 6 µm, fitting the human genome inside it implies a compression ratio of order 10,000 [3]. Strikingly, this level of packaging is achieved despite the large negative charge of the phosphate backbone — a lone DNA molecule would experience too much self repulsion to compact, this is overcome by salt screening and the inclusion of positively charged histone proteins. This combination of DNA and proteins is called chromatin; an illustration is show in figure 1.1a.

The fundamental chromatin unit is the 10-nm wide nucleosome: ∼147 base pairs of DNA in a left-handed super-helical turn around a core of eight histone proteins (two copies each of H4, H3, H2A and H2B) [5]. The X-ray structure of the nucleosome was resolved in 1997 with a resolution of 2.8 Å [6]. The current best resolved structure, Nucleosome Core Particle 147 (NCP147) PDB entry 1KX5, shown in figure 1.1b, has a resolution of 1.9 Å [4]. Successive nucleosomes are connected by linker DNA segments, which vary in length depending on the organism, cell type, and genomic region. Nucleosome Repeat Lengths (NRL = length of linker DNA + 147 bp) have been found to vary from 165 bp to 220 bp [7, 8]. In low salt concentration chromatin forms a "beads on a string" 10-nm fiber, as pictured in figure 1.1c top panel. Addition of physiological salt concentration (0.15 mol/L NaCl) and, optionally, the histone proteins (H1 linker histone), drives chromatin to condense into a more compact fiber, often referred to as the 30-nm fiber, as pictured in figure 1.1c lower panel.

The structure of chromatin, beyond the 10-nm fiber, remains an intense topic of research and debate [9–12]. The traditional textbook view (figure 1.1a) is that 10-nm chromatin folds into a regular and rigid 30-nm solenoid [13], zigzag [14–16], or heteromorphic [17] fiber. However, accumulating new evidence is now shifting

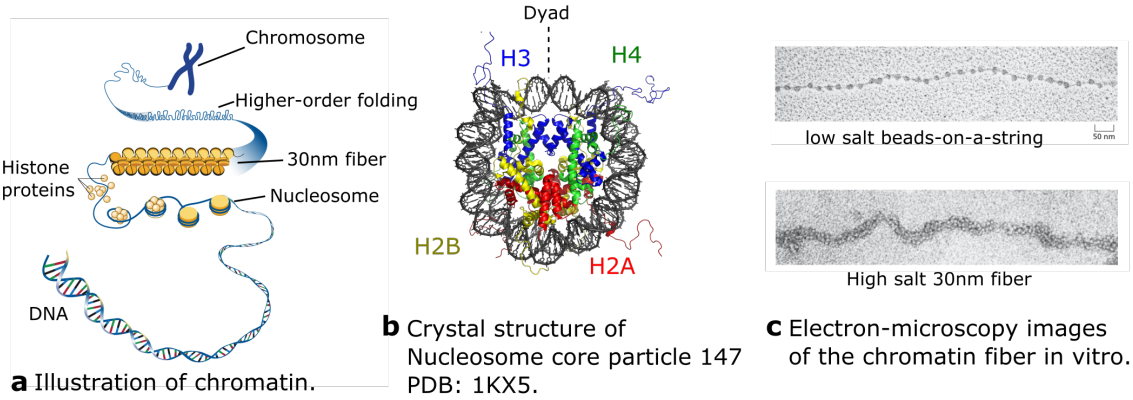

Figure 1.1: **Chromatin structure**. **(a)** "Text book" style illustration of chromatin: DNA wraps around histones forming nucleosomes, nucleosomes package together forming the chromatin fiber, higher order folding produces a metaphase chromosome. The image is taken from the National Human Genome Research Institute https://www.genome.gov/. **(b)** Crystal structure of a nucleosome, PDB entry 1KX5 [4]. The DNA is gray, the histone proteins are colored and labelled, there are 2 copies of each histone giving a total of 8 histone proteins forming the octamer core of the nucleosome. **(c)** Electron-microscopy images showing de-condensed "beads on a string" chromatin in the upper panel and a condensed 30nm chromatin fiber in the lower panel. These images are from Ref. [2].

the structural paradigm in favor of the 'liquid-like' or 'fluid-like' model [11, 18], which suggests that 10-nm chromatin fibers condense into an irregular and dynamic polymorphic ensemble [9, 19, 20]. The term liquid here is used to emphasize a compact chromatin structure that is absent of long-range translational order, and where nucleosomes can flow and relax easily. Several consistent models where chromatin exhibits a disordered organization based on the 10-nm fiber have been proposed recently, including hierarchical looping [21, 22], nucleosome clutches [23], multiplex higher order folding [24], and the sea of nucleosomes [25]. The liquid-like behavior of chromatin is consistent with its heterogeneity in vivo, e.g. varying DNA sequences, non-uniform NRLs, nucleosome free regions, and dynamic nucleosome sliding and breathing motions. Many of these factors can independently enable chromatin polymorphism (folding of fibers into irregular loops, hairpins, and bends) by giving rise to different nucleosome orientations and interactions [20, 26, 27].

Furthermore, it has been realized that chromatin and associated biomolecules can undergo liquid-liquid phase separation (LLPS) in vitro and in cells [28–40]. LLPS is postulated as a mechanism, alongside others [41], to explain genome compartmentalization without the use of physical membranes [28, 30, 31]. The emergence of intranuclear phase separation is intrinsically linked to the complex and crowded biomolecular environment of the cell nucleus [42, 43], i.e. the nucleoplasm is a multicomponent mixture of proteins and nucleic acids with varying compositions across different regions [44]. LLPS occurs when the energetics of the inter-molecular interactions overcome the entropy loss from demixing into the phase separated liquid states; this is controlled by the concentration and identities of the molecules in solution, the temperature, and the salt concentration [45, 46].

The advent of experimental techniques such as chromosome conformation capture [47–49], which creates genome wide interaction maps, and high resolution microscopy techniques such as ChromEMT [50], which can directly visualize in vivo

chromatin, have significant advanced the knowledge of chromatin structure. However, they still do not have sufficient resolution to resolve sub-nucleosome level interactions or dynamics; this is where computational modeling becomes an indispensable tool, allowing close up atomic level views and linking the gaps between nanometer scale process with larger scale processes observable in experiment.

In this work, we develop a multiscale chromatin model which incorporates information from all-atom nucleosome simulations, can represent chromatin at amino-acid and DNA base-pair resolution, and reach length scales of hundreds of nucleosomes and multiple interacting chromatin fibers. The model allows us to decipher how small structural nucleosome changes can effect higher order chromatin structure. The next two sections will cover some details on nucleosome structure and existing biomolecular coarse-grained models and chromatin-specific computational models.

1.1 Structural details of DNA and nucleosomes

Double stranded DNA (here referred to as simply DNA) is composed of two polynucleotide strands, strand 1 and strand 2, coiled around each other in a right-handed double helix as pictured in figure 1.2. A nucleotide consists of a nitrogenous base — Adenine (A), Guanine (G), Thymine (T), or Cytosine (C) — attached to a deoxyribose sugar and a phosphate group. Within a single strand of DNA, the phosphate of each nucleotide covalently binds to the deoxyribose sugar of the subsequent nucleotide, yielding the sugar-phosphate backbone. The idea of directionality of a strand of DNA emerges because the last atom of a strand can be either a 5' or 3' ribose carbon; hence, there are two possible directions in which one can read the sequence of bases that make of the strand: the direction from the 5' base to the 3' base, or from the 3' base to the 5' base. Since DNA is only synthesized in vivo in the 5' to 3' direction, this is defined as the forward direction.

DNA hydbrization — i.e., the pairing of two complementary strands to form a double stranded DNA — derives from the ability of bases to pair. That is, two complimentary bases (A is complimentary to T and C is complementary to G) can form long-lived canonical hydrogen bonds with one another. When two strands with complementary sequences of nucleotides hybridize, all bases of strand 1 pair via hydrogen bonds with the bases of strand 2, and subsequently, the base pairs establish $\pi - \pi$ stacking interactions. Because of the directionality of DNA, the two strands forming the hybridized DNA run in opposite directions and are, thus, referred to as anti-parallel. Within double stranded DNA, two sequential base-pairs (i.e. four bases) are called a base-pair step; as is customary, in this thesis we label a base-pair step as XY, where XY represents the two bases on the 5' to 3' strand of the DNA (or strand 1). The identity of the bases on strand 2 follows directly from this, as it corresponds to the complementary bases to X and Y. For example, base-pair step AC refers to bases AC on strand 1 and bases GT running in the opposite direction on strand 2. Due to the asymmetry of the nucleotides, the grooves of the double helix are different in size. The smaller groove is called the minor groove and the larger groove is termed the major groove.

The flexibility of DNA is greatest when bending in the direction of the major or minor grooves as opposed to bending in the direction of the phosphate backbone [51]. Therefore, when wound around the nucleosome core there are 10 bp (approximately

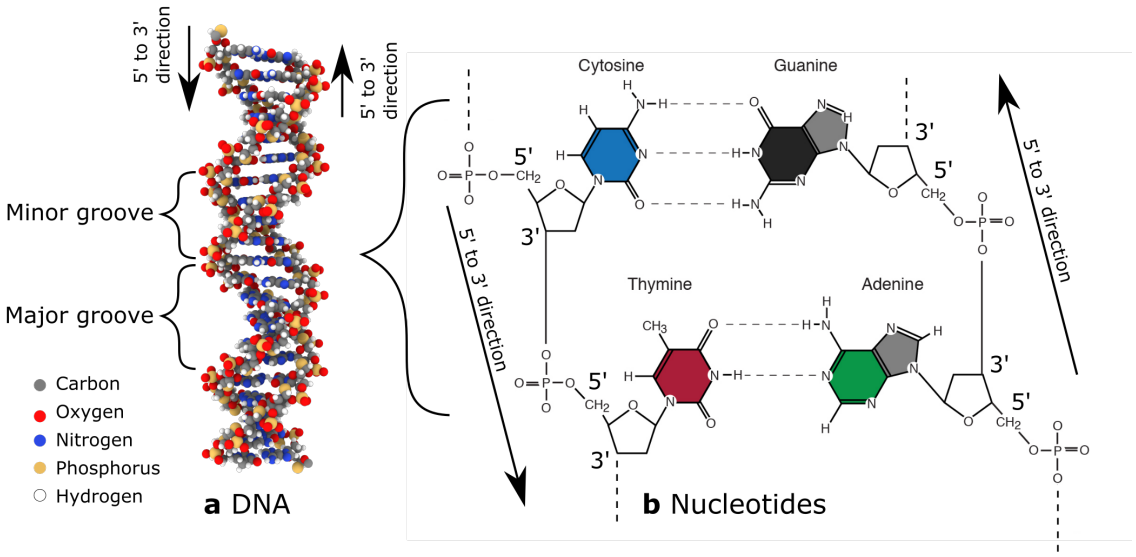

Figure 1.2: **DNA structure**. **(a)** Molecular structure of DNA, the atoms types are color coded corresponding to the legend. The major and minor grooves are indicated. **(b)** Complementary base-pairing of the four DNA nucleotides. This image is taken from the National Human Genome Research Institute https://www.genome.gov/.

the pitch of DNA) periodic oscillations in the degree of bending corresponding to where the major groove of the DNA faces inwards and again, with a 5 bp phase shift, when the minor groove faces inwards [5, 52].

It is well known that the mechanical properties of DNA are sequence dependent [53]; this, combined with the asymmetric bending of the nucleosomal DNA, led to the idea of the so-called nucleosome positioning code [54]. The nucleosome position code postulates that base-pair steps are not uniformly distributed along the nucleosome, but instead, some base-pair steps are more likely to be positioned at specific locations depending on their intrinsic flexibility. Experiments have shown that base-pair steps AA, AT, TT, and TA are more likely to be found at positions where the minor groove faces the histone protein core, and CC, CG, GC, and GG are more likely to be positioned at sites where the major groove faces the histone protein core [54–56]. Another prevalent DNA sequence seen in nature are poly-A tracts in nucleosome free regions [57, 58]. The repeated AA steps are stiff, which inhibits nucleosome formation [59].

The eight histones that make up the protein core of the nucleosome have a highly conserved structure comprised of a well-defined secondary structure, mostly alpha-helical, and terminal tail sections (the N and C terminal domains) which are Intrinsically Disordered Proteins (IDP). While the globular part of the histones make up the core of the nucleosome, the histone tails extend out from the nucleosome, mediating interactions with the nucleosomal DNA and other nucleosomes. The tails are common sites for post-translational modifications, and their flexibility and structure play important roles in nucleosome and chromatin structure. Indeed, by determining crystal structures of tail-less nucleosomes, Iwasaki et al [60] found that the removal of histone tails substantially decreases nucleosome stability.

The H1 linker histone is an additional highly abundant histone protein that does not form part of the nucleosome core. Instead, the H1 protein sits outside and

binds to the nucleosomal DNA at the so-called nucleosome dyad position. The dyad position is the location of the central base-pair in a typical 147 bp nucleosome and is labelled in figure 1.1b. While there are several variants of the H1 protein found in nature [61], all are comprised of an intrinsically disordered N-terminal tail of ∼24 residues, a highly structured globular domain of ∼80 residues, and an intrinsically disordered C-tail of ∼100 residues [62]. The linker histone has net positive charge, and hence, efficiently screens the electrostatic repulsion between the two DNA linker arms of a nucleosome. Such screening causes the DNA linker arms to change from an open configuration to a compact configuration, reducing their flexibility and bringing them together in a motif termed a 'DNA stem' [63]. The flexibility and unstructured nature of the long C-terminal domain of H1 is thought to be very important in the ability of linker histones to induce chromatin compaction, and modulate chromatin structure [63, 64]. The concentration of linker histones in cells has been found to range from 1 H1 per nucleosome to 0.1 per nucleosome, with a general trend that chromatin with longer NRLs has higher H1 concentrations [8].

1.2 Computational modeling of chromatin

DNA is a challenging molecule to simulate, as both the length scales and time scales relevant to describe its behavior in vivo span vast ranges. As previously mentioned, the total length of DNA that can be present in a nucleus is 2 m, while the distance between consecutive base pairs is only 3.4 Å. Another striking contrast is that while DNA sequence mutations are established on time-scales of years, the fastest electronic rearrangements within DNA take place on the sub-femtosecond scale [65]. Computational approaches to molecular simulation of DNA can be classified into three general areas; in ascending order of time and length scale, these are: Quantum mechanical, All-atom, and Coarse-grained descriptions. Although quantum mechanical methods are only feasible for a few DNA base-pairs, such methods are essential in the development of the force fields used in the next level up of all-atom models. All-atom methods are widely used in biomolecular modeling; in particular, they allow efficient sampling of DNA strands of up to ∼100 base pairs in explicit solvent and ions [65]. DNA molecules larger than ∼100 base pairs are prohibitive to simulate in part due to the cost of describing water molecules and ions explicitly — for a typical simulation box more than 90% of the atoms will be solvent. Indeed, implicit solvent descriptions can significantly increase the timescales and systems sizes achievable for atomistic DNA simulations (e.g. Pyne et al [66] combined explicit and implicit solvent molecular dynamics to simulate 339bp DNA mini-circles at all-atom resolution). Finally, we reach the level of Coarse-Grained (CG) models, where various atoms are grouped together into a coarse-grained bead, and usually (although not always [67, 68]) the solvent is treated implicitly. Models of low-resolution — i.e. where whole proteins or tens of DNA bases are represented by a single bead — are also called mesoscopic models [3]. For DNA, there is a wide-range of coarse-grained models with resolutions varying from 6 beads per base up to 10 thousand base-pairs per bead [65].

Some examples of high-resolution (i.e., where more than 1 bead is used to represent a base-pair) coarse-grained DNA models are: oxDNA [69], SIRAH [68], 3SPN [70], and MARTINI [71] (for a full review, see Dans et al [65]). A DNA model that uses 1 bead per base-pair is the rigid-base-pair model [72–76], which we

use in this work.

There is also a wide-range of coarse-grained models designed specifically to describe chromatin at varying resolutions. We note that the design of the coarse-grained models, and types of potentials to describe the interactions in each case is intricately linked to the method used for sampling (e.g. Monte Carlo versus molecular dynamics). A famous example is the model of Schlick et al [77]; this represents the nucleosome core (including DNA but excluding histone tails) using 300 discrete point charges, which positions and values are fitted to best approximate the full all-atom electric field around the full nucleosome at a specific distance. The flexible histone tails are modeled separately, using one bead per every five amino acids. The linker DNA is modeled using a modified version of the discrete worm-like chain model where each bead represents 10 base-pairs, with added negative charges [78]. In the original formulation, linker histones are represented by a three bead rigid body permanently attached to the nucleosome dyad. The model uses Monte-Carlo sampling with a salt dependent screened Coulomb interaction between all charged components. The model has been used for numerous investigations. Collepardo-Guevara and Schlick found that variations in the DNA linker length triggers chromatin polymorphism; that is combinations of different NRLs in the same fiber give rise to very different chromatin fiber shapes, such as bent ladders, hairpins, and loops [20]. Bascom and Schlick [79] simulated 100 nucleosome fibers and found that the overall chromatin compaction is dependent on a cooperation between the linker histone binding and the acetylation of the histone tails. Acetylation involves the addition of a negatively charged acetyl group to a positively charged lysine residue. This results in a folding of the tail reducing its interactions with other components [80].

The nucleosome model of Nordenskiöld's group [81] uses the following higher-resolution representation: one bead is used for each amino acid in the core protein, including histone tails, one bead is used for each base pair, and 4 beads describe the phosphates of a base-pair. The model also includes ions explicitly, and is sampled via molecular dynamics. Using this model, Nordenskiöld and colleagues investigated the effect of mono, di, and tri-valent cations on the intra and inter nucleosome interactions finding that increasing the cation charge causes the nucleosome-nucleosome interaction to switch from repulsive for monovalent cations to attractive for trivalent cations. Importantly, this model does not contain linker DNA, so the nucleosomes are not linked to one another but rather considered as separate composite particles.

The Monte Carlo model of the Olson group [82] pioneered the use of a rigid base-pair parameterization of the DNA combined with a charged bead (one per every three DNA base-pairs). The Olson model approximates the nucleosome core as a wedge shaped excluded volume object with the charged amino acids grouped into approximately 25 residues per charged bead. The histone tails are included as mobile point charges constrained within spherical regions. The interaction energies among beads consist of a screened Coulomb interaction. Collision detection is implemented to enforce excluded volumes. This model has been applied to compute chromatin fiber stiffness as a function of NRL. Additionally, the model is used as the basis of a lower-resolution coarse-grained model that represents each nucleosome with a single bead; this is done by parameterizing the nucleosome–nucleosome interactions using a rigid base-pair like potential.

The Schiessel group [83] created a single nucleosome model also using the DNA rigid base-pair model. They developed a Monte Carlo sampling approach to sample

the DNA sequence space — a technique they called Mutation Monte Carlo — and find the preferred base-pair sequences to form nucleosomes. Their results agreed with the nucleosome positioning code proposed by Segal et al [54] earlier. Fathizadeh et al used a molecular dynamics version of this model to investigate the binding and sliding of the nucleosomal DNA [84].

More recently, the de Pablo group created the 1CPN model which uses a multiscale approach [85]. At the first level, they have a high-resolution coarse-grained model for a single nucleosome based on the 3SPN DNA model combined with the AICG protein model [86] (a one site per amino-acid model) which has been used in previous single nucleosome studies [87, 88]. From this, by equating the interaction parameters with the free energy measurements, they then developed the second level lower-resolution coarse-grained chromatin model which has one cylinder per nucleosome (1CPN). The model can simulate the effects of different linker lengths and DNA sequences. It has also been used to study the free energy profiles of the center-to-center distance between two nucleosomes. The model gives good agreement with experimental salt-dependent chromatin sedimentation coefficients of 12-nucleosome 207 NRL chromatin.

The van Noort group used a rigid-base-pair resolution Monte Carlo chromatin model to investigate the behavior of chromatin unfolding by force [89, 90]. While accurate in the DNA description, the protein description in this model is oversimplified.

We have only mentioned a small number of existing models, which are most similar in scope and characteristics to the one we will develop in this thesis. A fuller review of the current sate of chromatin computational modeling, can be found in Refs. [3, 65, 91–93].

1.3 Thesis overview — multiscale methodology

In this work we develop a multiscale methodology for chromatin simulation (figure 1.3). At level 1 we have all-atom simulations of single nucleosomes; level 2 we have coarse-grained simulations at amino-acid and DNA base-pair resolution, with systems sizes of tens of nucleosomes; and level 3 we have simulations at nucleosome resolution with multiple chromatin fibers and hundreds of nucleosomes.

The level 1 all-atom simulations were performed in previous work by our research group and will not be discussed in detail [64].

The level 2 and level 3 CG models were developed in this work and will be explained in detail in the respective sections. Both use the "bottom-up" coarse-graining methodology, that is the level 2 model is created from the level 1 simulations and the level 3 model is created from the level 2 simulations. We have named the level 2 model the "chemically-specific model" because it retains the identity of the individual amino acids and DNA sequence and the level 3 model the "minimal model" as it is a minimal representation needed to represent the chromatin properties of interest.

Chapter 2 will cover some background theory on the topic of biomolecular simulation, chapter 3 will explain the development of the chemically-specific model, chapter 4 contains simulations done using the chemically-specific model, chapter 5 will explain the development of the minimal model, and chapter 6 will contain the simulations done using the minimal model.

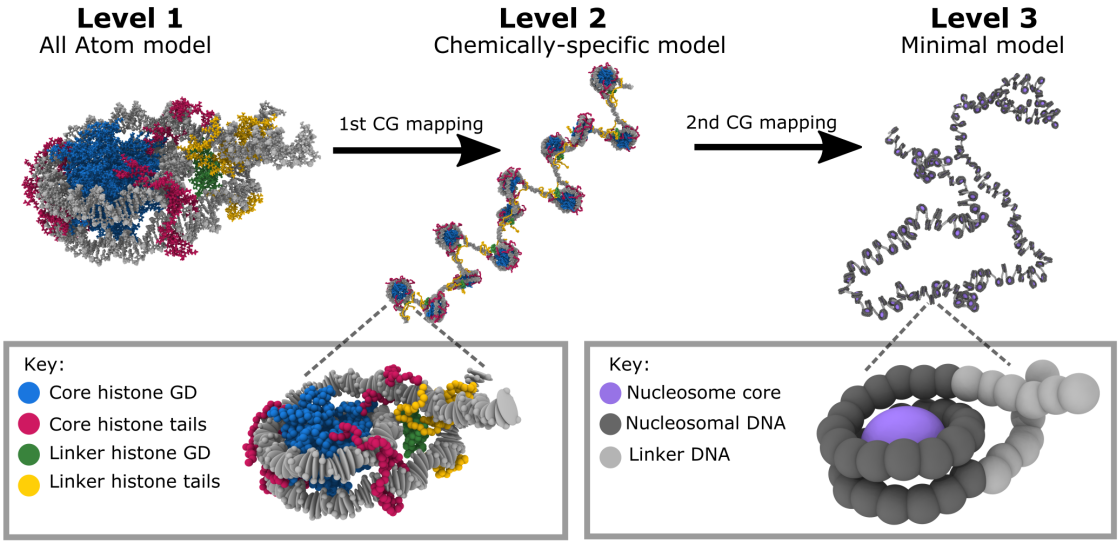

Figure 1.3: **Multiscale chromatin model**. **Level 1:** All-atom MD simulations of nucleosomes to obtain key structure and physical information. **Level 2:** Our chemically-specific coarse-grained model which represents DNA at base-pair resolution and protein at amino-acid resolution. This model is able to link elementary properties of nucleosomes to mesoscale behavior of oligonucleosomes. **Level 3:** Our minimal model, the nucleosome histone core is represented by a single bead and the DNA is described by one bead per 5 base-pairs. It is able to simulate chromatin fibers with hundreds of nucleosomes and perform coexistence simulations with over a thousand nucleosomes.

An animated illustration of our multiscale strategy can be found in https://sef43.gitlab.io/ along with a selection of videos related to simulations performed in this work.

1.4 Software used

Simulations were performed using the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) program [94] (version 3rd March 2020) with our custom code (see Code Availability). We used the program 3DNA [95] (version 2.3) within our model building methods. All data analysis was done using Python (version 3.8.5) with NumPy (version 1.19.2) and SciPy (version 1.5.2). All data was plotted using Matplotlib (version 3.3.2). Images were rendered using the Open Visualization Tool (OVITO) software [96] (version 3.0.0). We used the Weighted Histogram Analysis Method (WHAM) program [97] (version 2.0.9) to calculate PMFs.

1.5 Code availability

The source code associated with this work is available here: https://github.com/CollepardoLab/CollepardoLab_Chromatin_Model and https://doi.org/10.6084/m9.figshare.13663685.v1.

Chapter 2

Background theory — biomolecular simulation

In this chapter we go into the details of the theoretical and practical aspects of biomolecular computer simulations.

Contents

2.1 Statistical physics

We begin with a brief discussion of statistical physics, which is what our methods rely on, as it provides a framework for relating the microscopic properties of individual particles to the macroscopic properties of the entire system. A system is described by a set of position coordinates $\mathbf{r}$ and their conjugate momenta $\mathbf{p}$. The potential energy $E(\mathbf{r})$ of the system is a function of only the positions, the kinetic energy $K(\mathbf{p})$ is a function of only the momenta, and the total energy (the Hamiltonian H) is the sum of these $H(\mathbf{r}, \mathbf{p}) = E(\mathbf{r}) + K(\mathbf{p})$. Assuming the system is in contact with a heat bath of temperature T (i.e., it is in the canonical ensemble), we can write the probability density of finding the system in a state with energy $H(\mathbf{r}, \mathbf{p})$ as

$$p(H) = \frac{e^{-\beta H}}{Z}, \tag{2.1}$$

where Z is the partition function,

$$Z = \int e^{-\beta H} d\mathbf{r} d\mathbf{p}, \tag{2.2}$$

where $\beta = 1/(k_B T)$ is the inverse temperature and the integral is over the entire phase space, that is all the possible values of $\mathbf{r}$ and $\mathbf{p}$. For simplicity we have not included the commonly used normalization prefactor of $1/h^3$.

Macroscopic properties, or observables, are then calculated by averaging over the entire probability distribution,

$$\langle A \rangle = \int A(\mathbf{r}, \mathbf{p}) \frac{e^{-\beta H(\mathbf{r},\mathbf{p})}}{Z} d\mathbf{r} d\mathbf{p}, \tag{2.3}$$

where $A(\mathbf{r}, \mathbf{p})$ is the value of observable A when the system is in a state with specific values of $\mathbf{r}$ and $\mathbf{p}$. Thus, $\langle A \rangle$ is the expectation value, or measured macroscopic value, of A.

The form of the Hamiltonian implies that the potential and kinetic energy can be separated, and the probability distribution factored into a configurational $p_{\text{config}}(\mathbf{r})$ and a kinetic $p_{\text{kinetic}}(\mathbf{p})$ term, as follows:

$$p(H) = p_{\text{config}}(\mathbf{r}) \cdot p_{\text{kinetic}}(\mathbf{p}) = \frac{e^{-\beta E(\mathbf{r})} e^{-\beta K(\mathbf{p})}}{\int e^{-\beta E(\mathbf{r})} d\mathbf{r} \int e^{-\beta K(\mathbf{p})} d\mathbf{p}}. \tag{2.4}$$

Using the standard form of the kinetic energy of $K = \mathbf{p}^2/(2m)$ means that the kinetic energy distribution is simply a multivariate Gaussian, which can be integrated analytically:

$$p_{\text{kinetic}}(\mathbf{p}) = \frac{e^{-\beta \mathbf{p}^2/(2m)}}{\int e^{-\beta \mathbf{p}^2/(2m)} d\mathbf{p}} = (2\pi m k_B T)^{(-3/2)} e^{-\beta \mathbf{p}^2/(2m)}, \tag{2.5}$$

where the result is the so-called Maxwell-Boltzmann distribution. For many-particle systems, the configurational distribution,

$$p_{\text{config}}(\mathbf{r}) = \frac{e^{-\beta E(\mathbf{r})}}{\int e^{-\beta E(\mathbf{r})} d\mathbf{r}}, \tag{2.6}$$

cannot be calculated so easily, as it is usually a complicated function of the atomic coordinates. Configurational properties (such as bond lengths, angles, radius of gyration) depend only on the configuration part, i.e. the kinetic component vanishes when $d\mathbf{p}$ is integrated over:

$$p(E) = \int p(H) d\mathbf{p} = \frac{e^{-\beta E(\mathbf{r})}}{\int e^{-\beta E(\mathbf{r})} d\mathbf{r}} \frac{\int e^{-\beta K(\mathbf{p})} d\mathbf{p}}{\int e^{-\beta K(\mathbf{p})} d\mathbf{p}} = p_{\text{config}}(\mathbf{r}), \tag{2.7}$$

and observables for these values are calculated by integrating over the configurational probability distribution

$$\langle A \rangle = \int A(\mathbf{r}) \frac{e^{-\beta E(\mathbf{r})}}{Z} d\mathbf{r}. \tag{2.8}$$

In the typical approach of computational discretization, this could be computed by performing sums,

$$\langle A \rangle = \frac{\sum A(\mathbf{r}) e^{-\beta U(\mathbf{r})}}{\sum e^{-\beta(\mathbf{r})}}. \tag{2.9}$$

However, the number of elements in the sum is given by the number of grid points in discretized space (n) to the power of the dimension (d), times the number of particles (N), $n^{(dN)}$. For instance, for 100 particles in a 3d space discretized into 100 grid points, the number of elements in the sum already equals $100^{3\times100}$, which is intractable. Importantly, in many physical systems, the most probable configurations are strongly peaked around a small subset of all possible states. Hence, the

majority of points chosen by uniformly sampling phase space will give vanishingly small probabilities. If we instead choose points, not from a grid, but randomly from the Boltzmann distribution (which gives the weight of each state), the sum becomes

$$\langle A \rangle = \frac{1}{N} \sum_{i}^{N} A_i, \tag{2.10}$$

where A_i are the discrete values of A sampled from the Boltzmann distribution. This is known as importance sampling. Monte Carlo methods and appropriate molecular dynamics methods allow us to sample points in configurational space according to the correct Boltzmann weights.

2.2 Monte Carlo

The Monte Carlo method uses random sampling to obtain a sequence (or trajectory) of samples from the probability distribution of interest P. Importantly, we only require knowledge of a probability distribution f that is proportional to P. For our purposes, this is the Boltzmann distribution. If we have a system in state x, then we can compute the Boltzmann probability

$$f(x) = e^{-\beta E(x)}, \tag{2.11}$$

which is proportional to P,

$$P(x) = \frac{1}{Z} e^{-\beta E(x)}. \tag{2.12}$$

What we cannot easily compute is Z. If we generate a new trial system state x' which is perturbed from x, we wish to know if we should accept or reject this new configuration using an acceptance rule that guarantees x' is also sampled from the Boltzmann distribution. The transition probability is $p(x \to x')$. Once equilibrium is established, if we want to sample according to the correct Boltzmann weights, we require that the average number of moves from state x to state x' is balanced by the number of reverse moves from x' to x; this is the detailed balance condition, written as,

$$p(x \to x') f(x) = p(x' \to x) f(x'). \tag{2.13}$$

The transition probabilities can be separated into a generation and acceptance step,

$$p(x \to x') = g(x \to x') a(x \to x'). \tag{2.14}$$

The generation step can be chosen to be symmetric,

$$g(x \to x') = g(x' \to x). \tag{2.15}$$

For example, this could be a displacement by a random vector Δx. The acceptance probabilities can now be written as

$$\frac{a(x \to x')}{a(x' \to x)} = \frac{f(x')}{f(x)} = e^{-\beta(E(x') - E(x))}, \tag{2.16}$$

which is satisfied by choosing a as the Metropolis criterion,

$$a = \min(1, e^{-\beta(E(x') - E(x))}). \tag{2.17}$$

Note the ratio $f(x')/f(x) = P(x')/P(x)$, it does not matter that we cannot directly calculate Z.

A Metropolis Monte Carlo simulation has the following steps:

- Initialize: Start in state x
- Iterate:
  1. generate trial state x'
  2. compute $(E(x') - E(x))$ and the Metropolis criterion a.
  3. generate a uniform random number r between 0 and 1
     - if $r < a$: accept the move and set x to x'
     - else: keep the system in state x.
  4. append x to the trajectory (even if it stayed in the same state)

The trajectory of x values can now be used to compute the set of A_i in equation 2.10.

2.3 Molecular dynamics

Molecular dynamics samples the phase space of a system by solving the equations of motion. We start with a system in a state with particle coordinates $\mathbf{r}$, give them initial momentum $\mathbf{p} = m\mathbf{v}$ (where m is the mass and $\mathbf{v}$ the velocity), and propagate the system forward in time obeying Newton's equation of motion:

$$\mathbf{F} = m\frac{d^2\mathbf{r}}{dt^2} = -\boldsymbol{\nabla} E(\mathbf{r}). \tag{2.18}$$

This equation has to be solved numerically for all non-trivial systems. The most commonly used method is an explicit time-stepping scheme called the velocity-Verlet algorithm, a typical implementation is shown below:

$$\begin{aligned} v^{i+1/2} &= v^i + \frac{dt}{2m}F^i, \\ x^{i+1} &= x^i + dtv^{i+1/2}, \\ F^{i+1} &= -\nabla E(x^{i+1}), \\ v^{i+1} &= v^{i+1/2} + \frac{dt}{2m}F^{i+1}. \end{aligned} \tag{2.19}$$

Where x^i and v^i are the particle coordinate and velocity at timestep i, respectively, and dt is the discrete timestep. It can be shown that the global error of this method is of order dt^2 (second order), and that it is a sympletic integrator [98]. This means that although it will diverge from the "true" solution, as will all numerical schemes, it will oscillate around the true Hamiltonian ensuring that important quantities, such as energy and momentum, are conserved.

Molecular dynamics, as described so far, will sample the constant energy micro-canonical ensemble. To sample the canonical ensemble, the system needs to be connected to a heat bath (also called a thermostat). A variety of techniques exist for thermostating. For our purposes, we will focus on Langevin dynamics. This

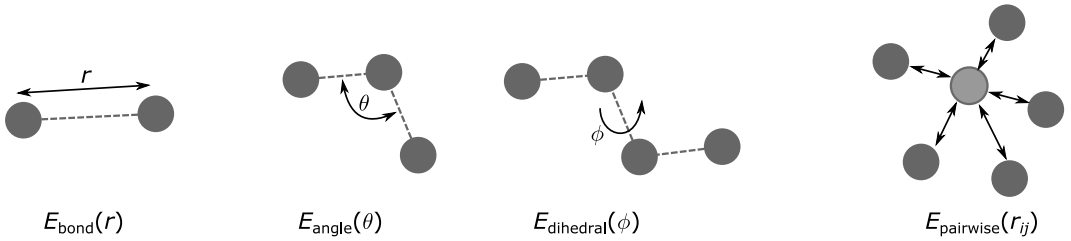

Figure 2.1: Typical force field terms.

involves the addition of a random force and a drag force to equation 2.18, which gives the Langevin equation:

$$m\frac{d^2\mathbf{r}}{dt^2} = -\boldsymbol{\nabla}E(\mathbf{r}) - \gamma\frac{d\mathbf{r}}{dt} + \sqrt{2\gamma k_B T}\mathbf{R}(t), \tag{2.20}$$

where γ is the friction constant and $\mathbf{R}$ is a vector of three delta correlated Gaussian random numbers, $\langle R(t)\rangle = 0$, $\langle R(t)R(t')\rangle = \delta(t-t')$. A numerical scheme for solving this equation is the Gronbech-Jensen Farago (GJF) integrator [99] shown below in a form suitable for direct implementation:

$$\begin{aligned} v^{i+1/2} &= v^i + \frac{dt}{2m}F^i, \\ x^{i+1} &= x^i + dtv^{i+1/2}, \\ F^{i+1} &= \frac{1}{1+\frac{\gamma dt}{2m}}\left[-\nabla E(x^{i+1}) - \gamma v^{i+1/2} + \sqrt{\frac{2\gamma k_B T}{dt}}\frac{(R^{i+1}+R^i)}{2}\right], \\ v^{i+1} &= v^{i+1/2} + \frac{dt}{2m}F^{i+1}. \end{aligned} \tag{2.21}$$

Note that the only differences between the implementation of the standard velocity-Verlet integrator and the Langevin integrator are the force calculation and the need to remember the random numbers used in the previous time-step.

2.4 Potentials and force fields

We have discussed methods to sample configurations with a potential $E(\mathbf{r})$ and now we will discuss how to represent and calculate E (and $\mathbf{F} = -\nabla E$). The potential energy of classical molecular systems is traditionally approximated as the sum of functions of bonds, angles, dihedrals, and pairwise interactions, as illustrated in figure 2.1.

2.4.1 Bonds

A bonded interaction is a function of the distance r between two particles with coordinates $\mathbf{r}_1$ and $\mathbf{r}_2$. A harmonic term is usually used,

$$E(r) = \frac{1}{2}k_b(r - r_0)^2, \tag{2.22}$$

where $r = |\mathbf{r}_1 - \mathbf{r}_2|$, k_b is the stiffness constant with units of energy per distance squared, and r_0 is the equilibrium bond length.

2.4.2 Angles

An angle interaction is a function of the angle θ created by three particles with coordinates $\mathbf{r}_1$, $\mathbf{r}_2$, and $\mathbf{r}_3$. We define two vectors $\mathbf{v}_1 = \mathbf{r}_1 - \mathbf{r}_2$ and $\mathbf{v}_2 = \mathbf{r}_3 - \mathbf{r}_2$, the angle is then calculated as $\theta = \cos^{-1}(\mathbf{v}_1 \cdot \mathbf{v}_2)/(|\mathbf{v}_1||\mathbf{v}_2|)$. A harmonic functional form is also commonly used,

$$E(\theta) = \frac{1}{2} k_\theta (\theta - \theta_0)^2, \tag{2.23}$$

where θ_0 is the equilibrium angle and k_θ is the stiffness with units of energy per angle squared. The range of θ is 0 to π.

2.4.3 Dihedrals

A dihedral interaction is a four body term that is a function of the dihedral (also called torsion) angle between the two planes created by the first three and last three sets of atoms respectively. If the displacement vectors between the 4 particles are $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$, then the normal vectors to the planes are $\mathbf{n}_1 = \mathbf{v}_1 \times \mathbf{v}_2$ and $\mathbf{n}_2 = \mathbf{v}_2 \times \mathbf{v}_3$, the dihedral angle ϕ is then given by

$$\cos(\phi) = \frac{\mathbf{n}_1 \cdot \mathbf{n}_2}{|\mathbf{n}_1||\mathbf{n}_2|}, \tag{2.24}$$

$$\sin(\phi) = \frac{\mathbf{v}_2}{|\mathbf{v}_2|} \cdot \frac{\mathbf{n}_1 \times \mathbf{n}_2}{|\mathbf{n}_1||\mathbf{n}_2|}, \tag{2.25}$$

where ϕ is in the range $-\pi$ to π. An example function for the potential is a cosine term,

$$E(\phi) = k_\phi \left[1 + \cos(n\phi)\right], \tag{2.26}$$

where n is an integer and k_ϕ is the energy term with units of energy per angle squared.

2.4.4 Pairwise

Pairwise interactions are functions of the distance between pairs of non-bonded particles, $r_{ij} = |r_i - r_j|$. An example of this is the Lennard-Jones potential

$$E(r) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right], \tag{2.27}$$

where ϵ is the depth of the potential and σ is the distance at which the potential is zero.

2.5 Coarse-graining

A very large amount of work has been done to create force fields for all-atom biomolecules with explicit water, e.g. Amber [100], CHARMM [101], GROMOS [102], OPLS [103]. However, despite the progress and high performance computing systems, it is not feasible to study large biomolecular systems, like chromatin, at the all atom level. Currently only single and di-nucleosome systems have been simulated

at meaningful timescales at full all-atom resolution; a comprehensive list of all-atom studies can be found in [104]. We must mention that a 427-nucleosome (1 billion atoms) all-atom simulation was performed in 2019 by Jung et al [105], but 130,000 cpu cores were needed to obtain a performance of just 1 ns/day.

A solution to facilitate the study of larger systems and/or reach longer timescales is to coarse-grain the system. In the same way that all-atom (classical) molecular dynamics is an approximation of the full quantum mechanical degrees of freedom, coarse-graining takes the fine-grained system (all-atom in our current analogy, but coarse-graining can be performed between any level of resolution) with coordinates $\mathbf{r}_\mathrm{f}$ and potential $E_\mathrm{f}(\mathbf{r}_\mathrm{f})$ and maps them, via a coordinate mapping function M^{coords} and potential mapping function $M^\mathrm{potential}$, to coarse-grained (CG) coordinates $\mathbf{R}_\mathrm{c}$ with potential $E_\mathrm{c}(\mathbf{R}_\mathrm{c})$,

$$\mathbf{R}_\mathrm{c} = M^\mathrm{coords}(\mathbf{r}_\mathrm{f}), \tag{2.28}$$

$$E_\mathrm{c} = M^\mathrm{potential}(E_\mathrm{f}), \tag{2.29}$$

where the subscripts 'f' and 'c' correspond to fine and coarse respectively. The idea is that observables will be approximately the same for both systems,

$$\langle O(\mathbf{R}_\mathrm{c})\rangle \approx \langle O(\mathbf{r}_\mathrm{f})\rangle, \tag{2.30}$$

while the number of degrees of freedom of the coarse-grained system is significantly smaller,

$$N_\mathrm{c} \ll N_\mathrm{f}. \tag{2.31}$$

Thus, with a fixed computational cost, a coarse-grained system can be simulated for longer timescales than its fine-grained counterpart.

To create the CG mapping one typically begins by deciding how to map the coordinates, e.g. place one CG coordinate (or CG bead) at the center of mass of each amino-acid, this is the coordinate mapping function M^coords. It is usually relatively easy to design simple and robust versions of M^coords. However, the challenge is to construct E_c such that equation 2.30 holds. While is it possible to write down a statistical physics definition of E_c in terms of the free energy [106]:

$$\exp[-\beta E_\mathrm{c}(\mathbf{R}_\mathrm{c})] = \int \delta(M^\mathrm{coords}(\mathbf{r}_\mathrm{f}) - \mathbf{R}_\mathrm{c}) \exp[-\beta E(\mathbf{r}_\mathrm{f})] d\mathbf{r}_\mathrm{f}, \tag{2.32}$$

it is not much use to the practitioner because, for any non-trivial system, it is not possible to calculate the integral as it involves integrating over all degrees of freedom of the fine-grained system. Therefore, in practice E_c must be created using parameter fitting techniques. This means a functional form of E_c is proposed, e.g. a Lennard-Jones interaction, then an observable to fit against is chosen, e.g. a radius of gyration, and simulations are run using the coarse-grained model. Subsequently, the CG observable is calculated and compared against the target value. The parameters of E_c are then optimized (using any sort of global minimization technique, e.g. grid search parameter sweep, gradient descent methods, simulated annealing etc).

Depending on what is being fitted, the CG methodology can be classed into two categories: 1. Physics-based/bottom-up — this is where properties from the fine-grain system (typically all-atom system) are directly fitted to; this could be the radial distribution functions using iterative Boltzmann inversion [107]. 2. Data-driven/knowledge-based/top-down — this is when the model is fitted to reproduce certain know experimental quantities, e.g. a persistence length of a polymer,

or known molecular structures such as protein-data bank crystallographic structures [108]. Of course many CG models utilize both methodologies [109].

When deriving a CG model, a compromise will have to be made between the accuracy, efficiency, and transferability. These terms respectively mean: how correctly does the model predict observables, how much have the degrees of freedom been reduced by upon coarse-graining, and is the model still correct when the domain is changed.

2.6 Enhanced sampling methods

Molecular dynamics of biomolecular systems, even when coarse-grained, is often limited by insufficient sampling. This is due to rough energy landscapes with many local minima separated by high-energy barriers, the system remains trapped in these local minima for long timescales. Enhanced sampling methods aim to overcome this limitation by driving the systems over the free energy barriers to more thoroughly sample the system phase space with less computational cost than conventional MD [110]. A variety of enhanced sampling methods have been developed, such as replica exchange [111], umbrella sampling [112], metadyanmics [113], and simulated annealing [114]. We will explain replica exchange and umbrella sampling in detail as we will use them in this work.

2.6.1 Replica exchange

Replica exchange, also known as parallel tempering, is an enhanced sampling technique that can be used on molecular simulations to improve sampling of the phase-space [115]. The idea is to run multiple replicas of the system in parallel, each with a different temperature, and periodically attempt to exchange which configurations are at which temperature. The higher temperatures allow the system to overcome free energy barriers and therefore improve the sampling. Using the Metropolis criteria to perform the exchanges ensures detailed balance is obeyed and that the lower temperature of interest is correctly sampling the Boltzmann distribution. The Metropolis exchange criteria for the probability to accept an exchange between replica 1 and replica 2 is

$$P(1 \leftrightarrow 2) = \min\left(1, \exp\left[\left(\frac{1}{k_B T_1} - \frac{1}{k_B T_2}\right)(E_1 - E_2)\right]\right), \tag{2.33}$$

where T_1, T_2 are the temperatures of replicas 1, 2; and E_1, E_2 are the current values of the potential energies for replicas 1 and 2 respectively. If molecular dynamics is being used (Replica-Exchange Molecular Dynamics REMD), then upon an acceptance the velocities should be rescaled by the change in temperature [111]:

$$\mathbf{v}_1' = \sqrt{T_2/T_1}\mathbf{v}_1, \tag{2.34}$$

$$\mathbf{v}_2' = \sqrt{T_1/T_2}\mathbf{v}_2. \tag{2.35}$$

The number of replicas needed scales as $\sqrt{N_f}$ where N_f is the number of degrees of freedom of the system [116].

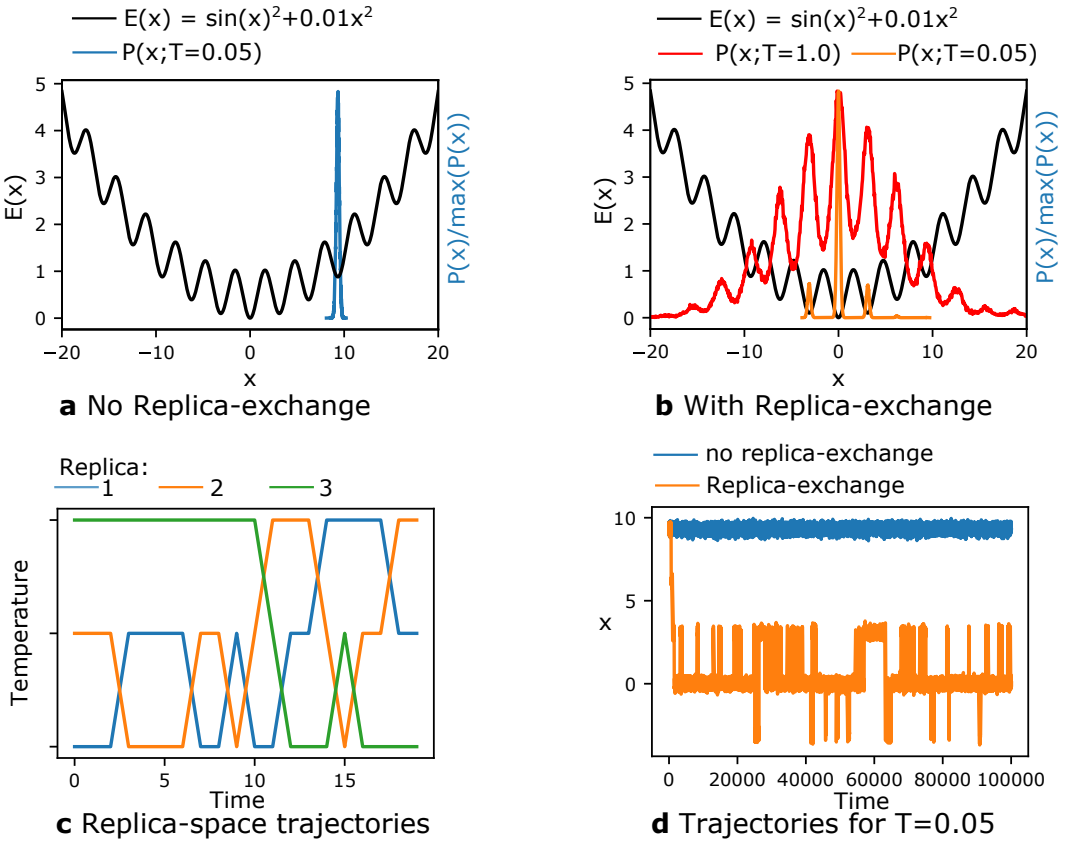

Figure 2.2: **Replica exchange method applied a toy system.** **(a)** A trajectory initialized at x=10, with a temperature of 0.05 (dimensionless units), will remained trapped in the meta-stable state of the rough potential energy surface $E(x) = \sin(x)^2 + 0.01x^2$. **(b)** Using replica exchange the sampled probability distribution at T=0.05 fully samples the global minimum, this is because the high temperature T=1 replica can freely move over the energy barriers. **(c)** Path of the trajectories in replica space. Line crossings correspond to accepted exchange attempts. **(d)** Comparison of the non-replica-exchange and replica-exchange trajectories, these are the trajectories used to plot the T=0.05 probability distributions in a and b.

Figure 2.2 illustrates the method applied to a toy system with a rough potential energy surface. Sub-figure 2.2a shows a trajectory initialized at x=10, with a temperature of 0.05 (dimensionless units), that stays trapped in a meta-stable state — it is unable to sample the global free energy minimum at x=0. In sub-figure 2.2b, using replica exchange with 3 replicas and a higher temperature of T=1, the T=0.05 trajectory now reaches the global minimum. We can see that T=1 is a high enough temperature that it can easily samples all states. The trajectories of the replicas are shown in sub-figure 2.2c and the trajectories of a system at T=0.05 initialized at x=10 is shown in sub-figure 2.2d, the blue line without replica exchange is trapped in a meta-stable state, the orange line using replica exchange effectively samples the global minimum.

Although replica exchange is traditionally performed using temperature as the exchange variable it is possible to exchange different Hamiltonians between replicas,

the general acceptance probability becomes [117],

$$P(1 \leftrightarrow 2) = \min\left(1, \exp\left[\frac{E_1(x_1) - E_1(x_2)}{k_B T_1} + \frac{E_2(x_2) - E_2(x_1)}{k_B T_2}\right]\right). \tag{2.36}$$

This can allow for more efficient schemes, e.g. solute tempering [118]. The key benefit of temperature replica exchange is that it is relatively simple to tune — by counting the degrees of freedom in the system one can work out the temperature scale that will give the desired acceptance ratio before actually running the simulation [116]. More complex bespoke Hamiltonian replica exchange methods usually need to be tuned manually by running trial simulations.

2.6.2 Umbrella sampling

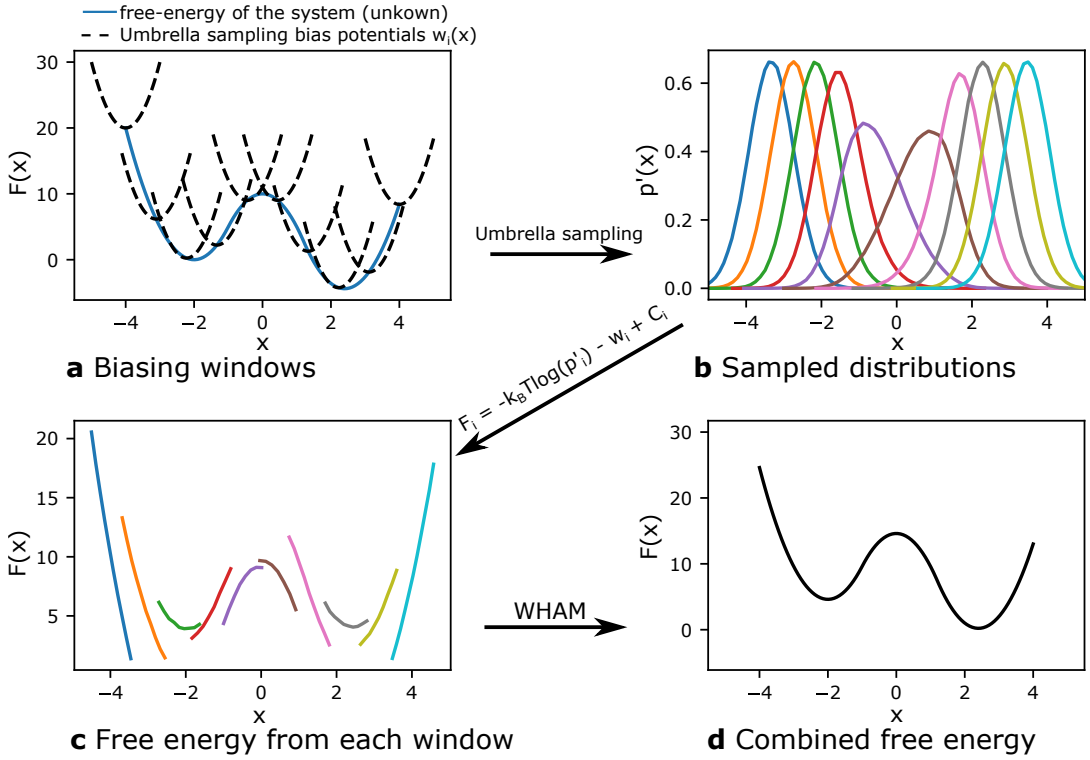

Figure 2.3: **Umbrella sampling method to compute a free energy profile. (a)** Multiple biasing potentials are placed across the collective variable x. The free energy of the system is the blue curve, in practice this is unknown. **(b)** Simulations are run for each w, these are called windows. The resulting biased probability distributions $P'(x)$ are plotted. **(c)** The unbiased free energies F_i from each window, they are each offset by a different C_i. **(d)** WHAM is used to combine the windows and compute the free energy curve.

The free energy of a system in the NVT ensemble is [98],

$$F = -k_B T \log(Z), \tag{2.37}$$

where Z is the canonical partition function. We have mentioned previously how in general Z is hard to calculate, however the most interesting information about a molecular system is given by the differences in the free energy across system states.

A reaction coordinate x (also called a collective variable) can be defined which is a continuous variable that distinguishes the different states. The simplest types of reaction coordinates are geometric distances, e.g. the distance between the center of mass of two molecules, but there are many possibilities. The collective variable x is a function of the atomic coordinates, $x(\mathbf{r})$, and multiple different realizations of $\mathbf{r}$ can map to the same x. The probability distribution of the system can be written in terms of x:

$$p(x) \propto \int e^{-\beta E(\mathbf{r})} \delta(x - x(\mathbf{r})) d\mathbf{r}, \tag{2.38}$$

where we have integrated the Boltzmann distribution over all degrees of freedom for each value of x. The probability distribution can be turned into a free energy [112]:

$$F(x) = -k_B T \log(p(x)) + C, \tag{2.39}$$

where C is a constant and unimportant as we are only interested in ΔF. In theory, the free energy profile along x could be computed by sampling the system in equilibrium and recording the probability histogram of the values of x which occur. However, for any non-trivial potential energy surface it will take a very long time to achieve sufficient sampling to get an adequately converged histogram — the high energy states will not be sampled.

To enhance the sampling a technique called umbrella sampling can be used. This adds additional biasing potentials $w(x)$ to the system to restrain it at certain values of x. The form of w is usually a harmonic term:

$$w(x) = \frac{1}{2} k (x - x_0)^2, \tag{2.40}$$

where k is a constant with units of energy per units of x squared. We then run multiple simulations, each restrained with a different w, and then combine them to generate a probability distribution that sufficiently samples the whole range of x. The process is illustrated in figure 2.3 and explained in more detail as follows.

With a biasing potential $w(x(\mathbf{r}))$ the potential energy becomes

$$E'(\mathbf{r}) = E(\mathbf{r}) + w(x(\mathbf{r})), \tag{2.41}$$

which leads to a probability distribution of

$$p'(x) \propto \int e^{-\beta(E(\mathbf{r}) + w(x(\mathbf{r})))} \delta(x - x(r)) d\mathbf{r} \propto p(x) e^{-\beta w(x)} \tag{2.42}$$

and a free energy of:

$$F'(x) = -k_B T \log(p(x)) + w(x) + C = F(x) + w(x) + C. \tag{2.43}$$

Thus the unbiased free energy F can be obtained by subtracting the biasing potential w from the biased free energy F'. However, when more than one biasing window is used the value of C cannot be neglected as it will be different for each window. For multiple biasing windows we have a set of unbiased (but offset by different C_i) free energies,

$$F_i(x) = -k_B T \log(p'_i) - w_i + C_i. \tag{2.44}$$

To compute F, or $P \propto e^{-\beta F}$, over the full range of x the different $F_i(x)$ need to be combined, this can be accomplished by the Weighted Histogram Analysis Method (WHAM) [112, 119, 120]. The WHAM equations are

$$P(x_j) = \frac{\sum_i^{N_w} h_i(x_j)}{\sum_i^{N_w} n_i e^{\beta(C_i - w_i(x_j))}}, \tag{2.45}$$

$$C_i = -k_B T \log \left(\sum_j P(x_j) e^{-\beta w_i(x_j)} \right), \tag{2.46}$$

where they have been written in a fully discretized form. $P(x_j)$ is the resulting unbiased probability distribution, j is the index for the discrete set of x_j that P is computed over. N_w is the number of windows, i is the index of each window, n_i is the number of data points (realizations of x) in the i-th window trajectory, $h_i(x_j)$ is the number of points in histogram bin j from trajectory i, and w_i are the biasing potentials. Both equations depend on each other so must be solved self-consistently. In practice this is solved iteratively: initial guesses of C_i are chosen, P is then calculated using equation 2.45, new values of C_i are then calculated using equation 2.46, and the process is iterated until the differences between successive values are sufficiently small.

The free energy curve along a collective variable is also called the Potential of Mean Force (PMF), in this work we will use the two terms interchangeably.

Chapter 3

Development of a chemically-specific chromatin model

In this chapter, I describe the original development of a chemically-specific coarse-grained model for chromatin (level 2 in our multiscale hierarchy). The development of this model is part of the publication [121].

Contents

3.1 Introduction

Inside cells, nucleosomes are highly heterogeneous both in terms of their chemical composition and structure. Importantly, such heterogeneity has been shown to sensitively affect the 3D organization of chromatin inside cells [122]. Thus, developing a chemically-specific model that can link the chemical heterogeneity and the spontaneous breathing motions of nucleosomes to chromatin self-assembly is highly desirable. Here, we embark on such a task, and devise a chemically-specific model for chromatin, where all proteins are resolved at single amino acid level and DNA at base-pair level. The sequence-dependent DNA and protein models are developed independently of each other. In the following sections, I describe each of these models, and their integration into our chromatin model.

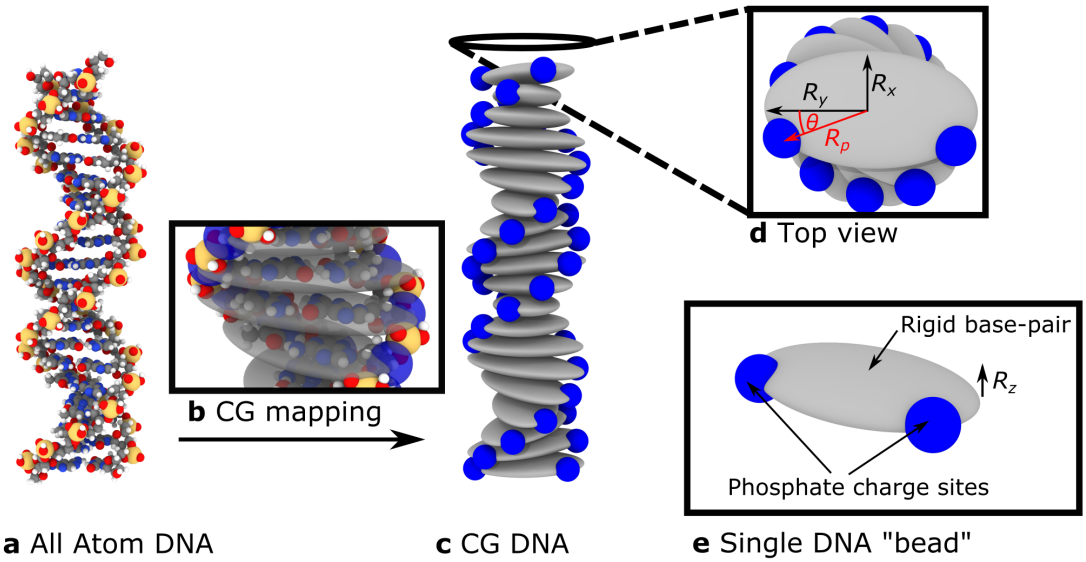

Figure 3.1: **Chemically-specific coarse-grained DNA model. (a)** All-atom structure of DNA. **(b)** Coarse-grained (CG) mapping procedure, illustrating how the base-pair shape can be approximated by an ellipsoid. **c** Our coarse-grained model of DNA, representing the all-atom structure in sub-figure a. **(d-e)** Zoomed in views of a single DNA bead with labels describing the geometry. The values of these parameters are in table 3.1.

3.2 Coarse-graining of DNA

DNA is a charged semi-flexible polymer with persistence lengths and torsional persistence lengths of the order of tens of base pairs [123]. These properties are challenging for standard bead-spring style polymer models to capture, and crucial to account for the correct organization of nucleosomes within chromatin. Therefore, we use the established rigid base-pair model of Olson and colleagues [124–129]. Importantly, because electrostatic interactions are well-established as major driving forces in chromatin organization (i.e., the balance between DNA–DNA electrostatic repulsion, DNA–histone tail anion-cation attraction, and the screening by counterions in solution), a critical step in our approach is to extend the rigid base-pair model by adding two charge sites per base-pair, centered at the position of the negatively charged phosphate groups. Hence, as pictured in figure 3.1, a DNA bead in our model is made up of three particles: one ellipsoid that represents the rigid base-pair and two point particles for the charge sites. These three particles are held rigidly together as a composite coarse-grained DNA bead. The ellipsoids are defined by a position vector, a quaternion orientation, and a shape (the length of the three ellipsoid axes). We set the mass of the ellipsoid to that of a DNA base-pair. The masses of the phosphate charge sites are set to a very small but non-zero value, so that they can be considered as virtual interaction sites. The non-zero mass is implemented simply to ensure computational stability. The ellipsoid shape and size, together with the relative position of the phosphate point particles within the ellipsoid, are set to approximate the geometry of an atomistic DNA base pair, via the parameters summarized in table 3.1.

The ionic effects of physiological monovalent ions in solution are approximated

Parameter	Value
R_x	5.5Å
R_y	10.0Å
R_z	1.75Å
R_p	8.5Å
θ_p	20°
Ellipsoid mass	650 g/mol
Phosphate mass	1e-6 g/mol
Ellipsoid charge	0
Phosphate charge	$-1e$

Table 3.1: Parameters of DNA model, corresponding to the labels in figure 3.1.

by an implicit solvent model via an screened Coulomb potential. The Langevin dynamics approximates thermal and viscous effects of the solvent.

3.2.1 Rigid base-pair model

The Rigid Base-Pair (RBP) model approximates DNA base-pairs as rigid planes, and bonded interactions between adjacent base-pairs are given by a six-dimensional harmonic potential which is a function of three displacements and three angles.

$$E_{\text{RBP}} = \frac{1}{2}\Delta\boldsymbol{\phi}\mathbf{K}\Delta\boldsymbol{\phi}^{\intercal} \tag{3.1}$$

$$\Delta\boldsymbol{\phi} = (\boldsymbol{\phi} - \boldsymbol{\phi}_0) \tag{3.2}$$

where $\boldsymbol{\phi} = (Dx, Dy, Dz, \tau, \rho, \omega)$ and $\mathbf{K}$ is a 6×6 stiffness matrix. $\boldsymbol{\phi}$ is the instantaneous value of the helical parameters, the components are shift, slide, rise, tilt, roll, and twist respectively and are pictured in figure 3.2. $\boldsymbol{\phi}_0$ are the mean helical parameter values analogous to equilibrium bond lengths. This potential is DNA sequence-dependent; therefore, each distinct base-pair step has its own values of $\mathbf{K}$ and $\boldsymbol{\phi}_0$. Considering the four different DNA bases, there are 16 possible base-pair steps; however, due to symmetry of the complementary base-pairing, this is reduced to only 10 unique steps: AA/TT, AC/GT, AG/CT, AT, CA/TG, CC/GG,CG,GA/TC,GC, and TA. The complementary steps (e.g AA and TT) differ only in the signs of shift and tilt which are flipped. These parameters have been defined in literature by X-ray crystallography [128] and molecular dynamics simulations [73, 75]. To calculate the helical parameters in our model we use the 'structure and conformation of helical nucleic acids: analysis program' (SCHNAaP) [130] algorithm which we describe in section A.2.

Traditionally, the RBP model has been used as a coarse-grained model to investigate DNA via Monte Carlo simulations, which only requires the potential energy and not its derivatives. However, to use the RBP potential in molecular dynamics simulations, the forces and torques must be defined. Following the method of Fathizadeh et al [84], the force in the k^{th} direction on a base-pair, due to the RBP potential E_{RBP} is,

$$F_k = -\frac{\partial E_{RBP}}{\partial r_k} = -\frac{\partial}{\partial r_k}\frac{1}{2}(\phi_i - \phi_{0i})K_{ij}(\phi_j - \phi_{0j}), \tag{3.3}$$

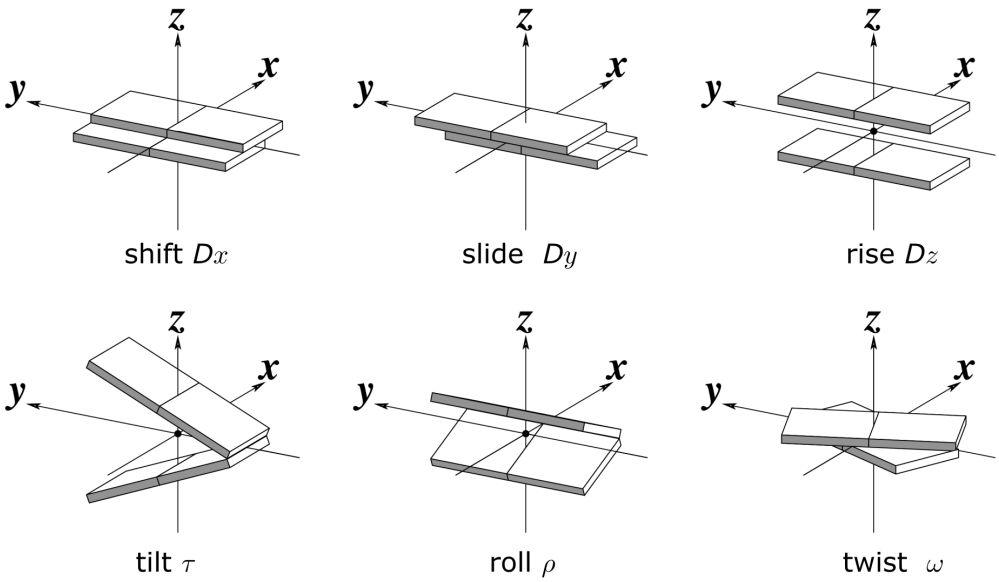

Figure 3.2: **DNA base-pair step helical parameters.** All diagrams show a positive value of the respective parameter. The shaded side denotes the minor groove side of the base-pairs. The axis are defined such that the x axis points from the minor-groove side to major-groove side, the z axis points in the forward direction of strand 1 (strand 1 is the left side if the minor groove is at the front), and the y axis is therefore the cross product of z and x. These images were reproducing using 3DNA [131, 132].

where using the symmetry of $\mathbf{K}$ gives,

$$F_k = -\frac{\partial \phi_i}{\partial r_k} K_{ij}(\phi_j - \phi_{0j}). \tag{3.4}$$

The three rotational components of the partial derivative of $\boldsymbol{\phi}$ with respect to r_k vanish because tilt, roll, and twist do not change on translation of the base-pairs.

$$\frac{\partial \phi_i}{\partial r_k} = 0, \quad i = 4, 5, 6. \tag{3.5}$$

To compute the partial derivative of the spatial components, we note that the calculation method (see section A.2) of shift, slide, and rise uses the equation:

$$\phi_i = T_{ij} x_j, \quad i = 1, 2, 3, \tag{3.6}$$

where T_{ij} are the components of the mid step orientation matrix $\mathbf{T}$, and x_j are the components of the displacement vector between the two base-pairs. Taking the partial derivative gives:

$$\frac{\partial \phi_i}{\partial r_k} = T_{ij} \frac{\partial x_j}{\partial r_k} = T_{ij} \delta_{jk} = T_{ik}, \tag{3.7}$$

where we note that $\mathbf{T}$ does not vary with ∂r_k. Therefore, the force is given by:

$$F_k = -T_{ik} K_{ij} \Delta \phi_j, \tag{3.8}$$

for $k = 1, 2, 3$, where i sums from 1 to 3 and j sums from 1 to 6.

The torque $\boldsymbol{\tau}$ is given in component form by:

$$\tau_k = -\frac{\partial E}{\partial \theta_k} \tag{3.9}$$

Where $\partial\theta_k$ is an infinitesimal rotation about the k^{th} axis. Unlike the force, this cannot be computed analytically, so we approximate it as a central finite difference:

$$\tau_k = -\frac{E(+\Delta\theta_k) - E(-\Delta\theta_k)}{2\Delta\theta}, \tag{3.10}$$

where $\Delta\theta_k$ represents rotating the base-pair by small angle of magnitude $\Delta\theta$ about the k axis. In practice, we use a value of 0.00001 radians for $\Delta\theta$. The reason why it cannot be computed analytically is because the function that computes the helical parameters is not a closed form analytical equation, but rather a heuristic computational procedure (given in section A.2).

3.2.2 Pairwise terms

To model the electrostatic interaction between DNA we use a screened Coloumb interaction which approximates the screening effects of counterions in solution. This has the following form:

$$E_{\text{Electrostatic}} = \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r} e^{-\kappa r}, \tag{3.11}$$

where q_i, q_j are the charges of the interacting pair of particles separated by distance r, ϵ_0 is the vacuum permittivity, ϵ_r is the relative permittivity (which we set to 80 corresponding to water for the entirety of this work), and κ is the inverse Deybe length λ_D, which is a function of the chosen monovalent salt concentration c.

$$\kappa^{-1} = \lambda_D = \sqrt{\frac{\epsilon_0\epsilon_r k_B T}{2 \times 10^3 N_A e^2 c}}; \tag{3.12}$$

where k_B is the Boltzmann constant, T is the temperature, N_A is the Avogadro constant, e is the elementary charge, and c is the monovalent salt concentration in units of mol/L.

This interaction is only counted between DNA beads that are not directly bonded. For clarity, as pictured in figure 3.3, this means that a DNA strand of three base-pairs in length has the following interactions: The two phosphates on base-pair 1 each interact with the two phosphates on base-pair 3. The phosphates on base-pair 2 do not interact with anything. The ellipsoids have no pairwise interaction because their charge is zero.

Typically, additional excluded volume terms, usually in the form of Lennard-Jones style interactions, would be used to prevent molecules from overlapping with each other. However, when the system contains only DNA, we have found that at the relatively low salt concentrations we use in this work (<0.3M), the electrostatic repulsion alone is sufficient to account for DNA–DNA excluded volume. When we introduce the protein model, additional excluded volume terms will be included between DNA and proteins.

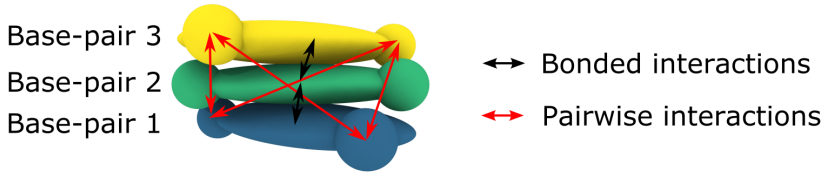

Figure 3.3: **DNA model interactions**. The bonded interactions (RBP potential) occur between the adjacent ellipsoids, the pairwise electrostatic interactions occur between phosphates on non-directly bonded base-pairs.

3.2.3 DNA energy function

To summarize the total potential energy function of the chemically-specific CG DNA model is:

$$E = \sum_{\text{DNA-bonds}} E_{\text{RBP}} + \sum_i \sum_{j<i} E_{\text{Electrostatic}}, \tag{3.13}$$

$$E_{\text{RBP}} = \frac{1}{2}\Delta\boldsymbol{\phi}\mathbf{K}\Delta\boldsymbol{\phi}^{\intercal},$$

$$\Delta\boldsymbol{\phi} = (\boldsymbol{\phi} - \boldsymbol{\phi}_0),$$

$$E_{\text{Electrostatic}} = \frac{q_i q_i}{4\pi\epsilon_0\epsilon_r r} e^{-\kappa r}.$$

3.3 Coarse-graining of proteins

For our CG protein model we use the model of Dignon et al [133], which represents each amino-acid residue with a single point-particle bead that has a sequence-dependent set of parameters, including charge, excluded volume radius, hydrophobicity, and mass (table 3.2). As pictured in figure 3.4, the beads are positioned on the C_α atoms of the amino acid they represent. Furthermore, amino-acids are categorised into either Intrinsically Disordered Protein (IDP) domains which are fully flexible and have a bond topology that follows the protein backbone, or globular domains which are regions that maintain secondary structure which is enforced in the model with a elastic network model.

3.3.1 Bonded interactions

The bonded potential for IDPs is a harmonic interaction of the form

$$E_{\text{bonds}} = \frac{1}{2}k\left(r - r_0\right)^2, \tag{3.14}$$

where k is the bond energy, r is the current bond length and r_0 is the equilibrium bond length. For all bonds, we use the same parameters. That is, $k = 10\,\text{kcal}\,\text{mol}^{-1}\,\text{Å}^{-2}$and $r_0 = 3.5\,\text{Å}$, as proposed by Dignon et al [133].

The globular domains are kept restrained in their fixed secondary structure by the use of an elastic network model, specifically a Gaussian Elastic Network model (GNM)[134]. Using the reference all-atom protein structure, all beads in the globular domain regions which are within 7.5 Å of each other are bonded together using a

Amino acid	Mass (g/mol)	σ (Å)	Charge (e)
ALA	71.08	5.04	0
ARG	156.2	6.56	1
ASN	114.1	5.68	0
ASP	115.1	5.58	-1
CYS	103.1	5.48	0
GLN	128.1	6.02	0
GLU	129.1	5.92	-1
GLY	57.05	4.50	0
HIS	137.1	6.08	0.5
ILE	113.2	6.18	0
LEU	113.2	6.18	0
LYS	128.2	6.36	1
MET	131.2	6.18	0
PHE	147.2	6.36	0
PRO	97.12	5.56	0
SER	87.08	5.18	0
THR	101.1	5.62	0
TRP	186.2	6.78	0
TYR	163.2	6.46	0
VAL	99.07	5.86	0

Table 3.2: Amino acid parameters taken from Ref. [133].

harmonic interaction term in the same form as equation 3.14. The value of k is also set to $10\,\mathrm{kcal/mol/Å^2}$, and the values of r_0 are set equal to the value in the reference structure.

3.3.2 Pairwise interactions

Pairwise interactions involve electrostatics in the form of a screened Coulomb potential

$$E_{\text{Electrostatic}} = \frac{q_i q_i}{4\pi\epsilon_0\epsilon_r r} e^{-\kappa r}, \tag{3.15}$$

where the parameters are the same as for equation 3.11. Additionally, for short range interactions we use the experimentally parameterized Kim-Hummer model [135], which provides both excluded volume effects and amino-acid specific attraction/repulsion approximating hydrophobic effects. This takes the form of a shifted and scaled Lennard-Jones potential:

$$E_{\text{KH}} = \begin{cases} E_{\text{LJ}} + (1-\lambda)\epsilon, & \text{if } r \leq 2^{1/6}\sigma, \\ \lambda E_{\text{LJ}}, & \text{otherwise}\ , \end{cases} \tag{3.16}$$

$$E_{\text{LJ}} = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]. \tag{3.17}$$

Where each amino acid pair has unique values of ϵ, λ, and σ which are determined from

$$\epsilon = |\alpha(\epsilon_{MJ} - \epsilon_0)| \tag{3.18}$$

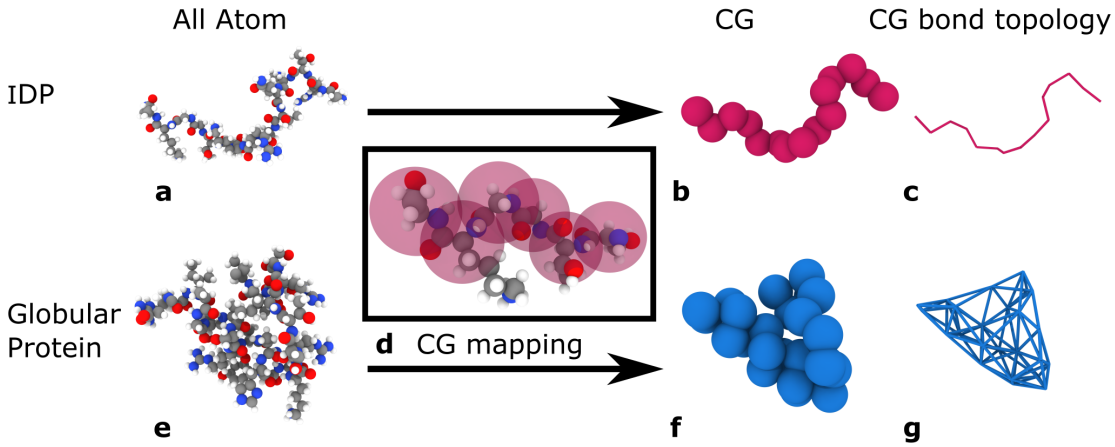

Figure 3.4: **Coarse-graining of protein**. **(a)** Example All-atom structure of a Intrinsically Disordered Protein (IDP). **(b)** Coarse-Grained (CG) structure representing (a). **(c)** Bond topology of the IDP corresponding to (b). **(d)** Illustration of the CG mapping procedure, the CG beads are positioned on the C_α atoms of the all-atom amino-acids. **(e)** Example all-atom structure of a globular protein, **(f)** CG structure representing (e), **(g)** Elastic network bond topology of the globular protein, corresponding to (f).

$$\lambda = \begin{cases} 1, & \text{if } \epsilon_{MJ} \leq \epsilon_0, \\ -1, & \text{else}, \end{cases} \tag{3.19}$$

$$\sigma = \frac{1}{2}(\sigma_i + \sigma_j), \tag{3.20}$$

where ϵ is from the Miyazawa-Jerningan statistical contact potential [136] (each amino acid pair has its own value), ϵ_0 and α are constants, which have six possible options identified in the original literature. Following Dignon et al [133], we use parameter set D for interactions between IDPs and set A for when either amino-acid in the interacting pair is part of the globular domain. Both sets are listed in tables D.4 and D.5. The σ_i values are the exclude volume sizes of the amino acid and are in table 3.2. The potential is plotted in figure 3.5 for a range of values.

The pairwise interactions are turned off for beads that are connected by a bond. Furthermore, a bead that is part of a globular domain GNM will not have pairwise interactions between any beads in the same GNM. The GNM restrains the overall structure, internal pairwise terms will have no effect so can be dropped to reduce computation time.

3.4 Chromatin model

To simulate chromatin we combine the DNA model and the protein model creating our 'chemically-specific' CG chromatin model (figure 3.6), this is readily doable as both models are at similar resolutions and share the same electrostatic potential. The only addition needed is to generate parameters for the E_{KH} interaction between DNA and proteins.

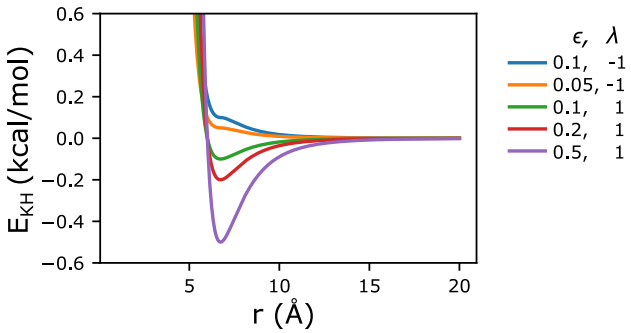

Figure 3.5: **The KH potential (equation 3.16.)** Plotted for the range of values typical of the model.

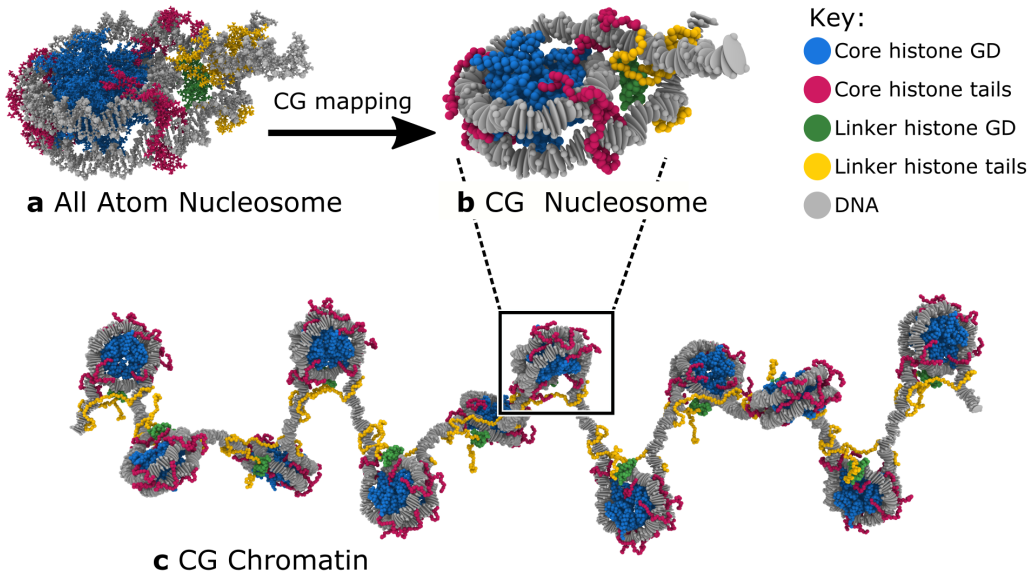

Figure 3.6: **Chemically-specific coarse-grained chromatin model (Level 2 in the multi-scale method hierarchy).** **(a)** All-atom nucleosome ∼30,000 atoms not including solvent. **(b)** Coarse-grained (CG) nucleosome ∼1000 beads, with implicit solvent. **(c)** CG 12-nucleosome chromatin.

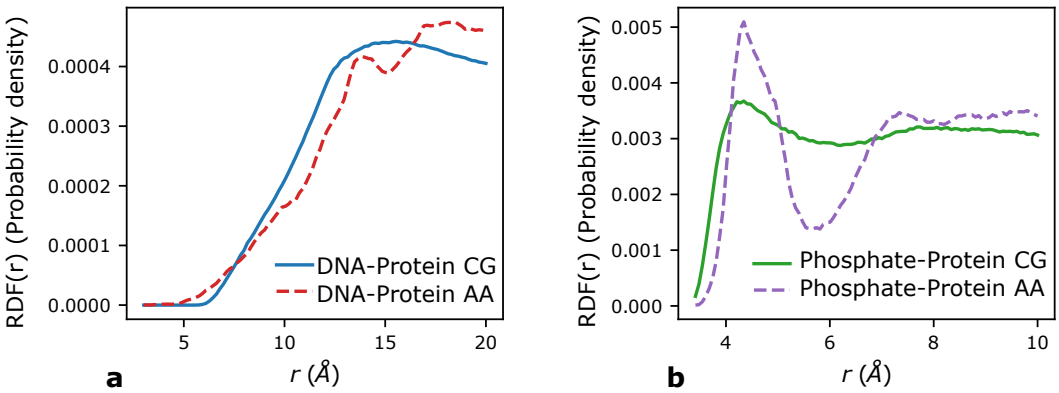

Figure 3.7: Comparisons of the All-Atom (AA) and coarse-grained (CG) radial distribution functions (RDFs) for **(a)** DNA ellipsoid–protein and **(b)** DNA phosphate–protein.

3.4.1 Fitting DNA-protein interaction

To parameterize the hydrophobic interaction between DNA and the different amino acids — i.e., the parameters for the potential E_{KH} between DNA and protein beads — we optimize the DNA ellipsoid–protein and the Phosphate–protein radial pair-wise distance distribution functions (RDF) of the coarse-grained simulations to match those computed from our 211-bp all-atom simulations of single nucleosomes [64]. We compute the RDF as:

$$\text{RDF}(r) = \frac{n_r}{4\pi r^2 \Delta r}, \tag{3.21}$$

where n_r is the number of particles found at distance r in a spherical shell of thickness Δr. To find n_r, we construct a histogram of all pair distances where n_r are the bin heights, and Δr are the bin widths. By plotting the RDFs from the all-atom simulations we first estimated $\sigma_{\text{DNA-protein}}$ and $\sigma_{\text{Phosphate-protein}}$ by reading off the location of the maximum of the RDFs. We then optimized the corresponding ϵ values by performing a grid search, i.e. running the CG nucleosome model for different trial ϵ values, computing the RDF and comparing to the all-atom RDF. For simplicity, we treat all DNA-protein interactions as being the same, note that the protein beads still have their unique value of intrinsic charge; thus, the DNA-protein pair potential is amino acid sequence-dependent due to the electrostatic contribution. The resulting fitted values are given in table 3.3, and the RDFs are plotted in figure 3.7.

Interaction	σ (Å)	ϵ (kcal/mol)	λ
DNA ellipsoid–protein	8	0.01	1
DNA phosphate–protein	4	0.1	1

Table 3.3: DNA–protein interaction parameters for E_{KH}.

For DNA-DNA interactions the E_{KH} term is set to zero, the DNA-DNA electrostatic repulsion is sufficient to account for the excluded volume.

3.4.2 Total chromatin energy function

The total potential energy function of our chromatin model is:

$$E = \sum_{\text{Protein-bonds}} E_{\text{Bonds}} + \sum_{\text{DNA-bonds}} E_{\text{RBP}} + \sum_{i}\sum_{j<i} E_{\text{Electrostatic}} + \sum_{i}\sum_{j<i} E_{\text{KH}}. \quad (3.22)$$

where

$$E_{\text{bonds}} = \frac{1}{2}k\left(r - r_0\right)^2, \quad (3.23)$$

$$E_{\text{RBP}} = \frac{1}{2}\Delta\boldsymbol{\phi}\mathbf{K}\Delta\boldsymbol{\phi}^{\intercal} \quad (3.24)$$

$$\Delta\boldsymbol{\phi} = (\boldsymbol{\phi} - \boldsymbol{\phi}_0) \quad (3.25)$$

$$E_{\text{Electrostatic}} = \frac{q_i q_i}{4\pi\epsilon_0\epsilon_r r}e^{-\kappa r}, \quad (3.26)$$

$$E_{\text{KH}} = \begin{cases} E_{\text{LJ}} + (1-\lambda)\epsilon, & \text{if } r \leq 2^{1/6}\sigma, \\ \lambda E_{\text{LJ}}, & \text{else }, \end{cases} \quad (3.27)$$

$$E_{\text{LJ}} = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]. \quad (3.28)$$

The parameter definitions and values for all terms can be found in the preceding sections and a summary is given in table D.2.

3.5 Breathing and non-breathing nucleosomes

We developed two versions of the model: (a) breathing (i.e., with nucleosomes that are allowed to breath spontaneously as found in vivo) and (b) non-breathing (i.e., with nucleosomes that are constrained to remain fully wrapped). We expand on the importance of these differences in the next Chapter. In practice, these two versions differ only in how the DNA beads are bound to the histone protein core. In the breathing case, DNA ellipsoids and amino acid beads interact exclusively via the pairwise interactions $E_{\text{Electrostatic}}$ and E_{KH} — this leaves the DNA free to bind and unbind spontaneously due to thermal fluctuations (i.e., "breathe"), and slide around the nucleosome core. In contrast, non-breathing simulations further constrain nucleosomal DNA by permanently bonding it to the histone core; this is implemented by including the DNA beads in the GNM with the same 7.5 Å threshold. This prevents these nucleosomes from breathing and sliding, and hence, forces them to remain fully wrapped.

3.6 Generating initial structures

Our reference structure for generating the CG chromatin structures is the all-atom nucleosome structure we use in our previous work [64]. This is a 211 bp nucleosome (modelled based on PDB structure 1KX5 [4]) with one H1 linker histone protein added to the nucleosome dyad.. The all-atom simulations were run using Bias exchange Meta-dynamics simulations, and the configuration of the most populated cluster was taken as our reference structure.

Our method for creating a CG chromatin fiber from a single nucleosome structure consists of first mapping the all-atom nucleosome into our CG representation. Subsequently, we replicate the CG nucleosome and join multiple nucleosomes together following the DNA polymer.

Firstly, from the all-atom structure we fit rigid base pairs (position vector and an orientation matrix) to the DNA using the software x3DNA [131]. The positions of all protein beads are defined by extracting the coordinates of the C-α atom for each amino-acid. The protein beads are categorised, using the definitions in table 3.4, into globular domains or histone tails (IDPs).

Histone	Tail region residues
H3	1-40
H4	1-25
H2A	1-20, 114-128
H2B	1-25
H1	1-24, 95-194

Table 3.4: Definition of histone tail regions. For H2-H4 all other residues are classified as belonging to the central globular region — the histone core. For H1 the other residues belong to the H1 linker histone globular domain, which is separate and distinct from the core globular domain.

In total, the nucleosome core has ten disordered tails and H1 has two. The nucleosome core globular domain amino-acids are all incorporated into one GNM. The H1 GD is incorporated into its own separate GNM. To create the GNMs, we create bonds between all pairs of protein beads that are within 7.5Å of each other. The DNA strand is then created by connecting beads in the order of the DNA sequence. Similarly, histone tail bonds follow the protein backbones.

This gives rise to the structure of a 211 bp single nucleosome; that is, a list of coordinates for all beads $\mathbf{x}_i$, a list of quaternions for the DNA ellipsoids $\underline{q}_i$, and a bond topology. Depending on the desired nucleosome repeat length, DNA bases are removed equally from each end of the DNA sequence.

To create a chromatin fiber of N nucleosomes, we repeatedly replicate the nucleosome and append it to the the current chromatin fiber, following the DNA strand. To position it in the correct location, that is a location which approximately minimizes the DNA bond energy, we first need to calculate the equilibrium position and orientation for the $n+1$ DNA bead, where n is the current number of DNA beads.

If $\mathbf{x}_n$ is the coordinate of the current terminal DNA bead, and $\underline{q}_n$ is the orientation, we compute $\mathbf{x}_{n+1}$ and $\underline{q}_{n+1}$ as

$$\mathbf{x}_{n+1} = \mathbf{x}_n + D_z \cdot \hat{\mathbf{z}}_n, \tag{3.29}$$

$$\underline{q}_{n+1} = rotate(\underline{q}_{end}, \hat{\mathbf{z}}_n, \omega), \tag{3.30}$$

where D_z is the equilibrium value of DNA rise, ω the equilibrium value of DNA twist, $\hat{\mathbf{z}}_n$ is the z direction unit vector of DNA bead n, and $rotate(\underline{q}, \mathbf{r}, \theta)$ is a function which rotates $\underline{q}$ by angle θ about axis $\mathbf{r}$.

We then take the coordinates and orientation of the first DNA bead in the reference nucleosome $\mathbf{x}_1$ and $\underline{q}_1$, and, changing the representation of $\underline{q}_1$ into matrix

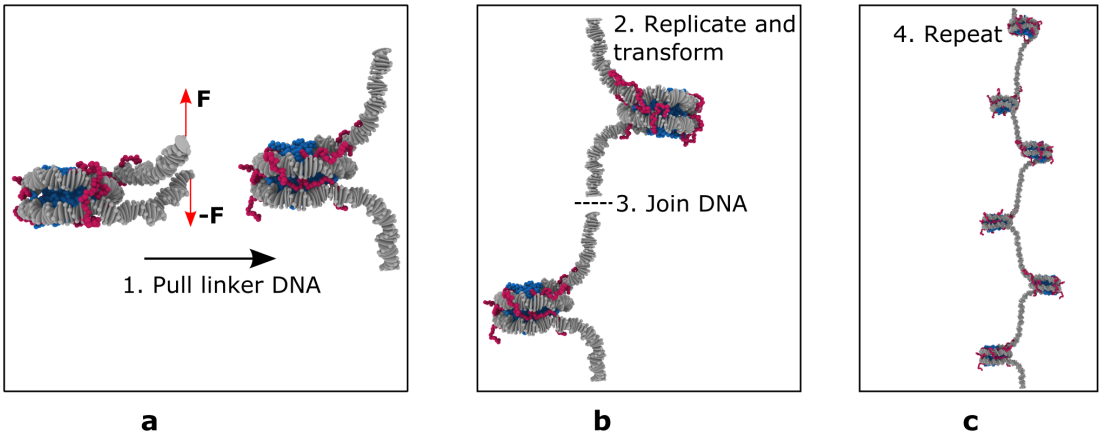

Figure 3.8: **Creating chromatin arrays from nucleosomes. (a)** The linker DNA arms are pulled in opposite directions. **(b)** The nucleosome is replicated and transformed into a position where joining the DNA results in a DNA bond that is close to its energy minima. **(c)** A "A beads on a string" chromatin array.

$\mathbf{R}$ (see equation A.5 in appendix A.1), we construct a 4x4 matrix with the entries as:

$$\mathbf{M}_1 = \begin{bmatrix} R_{11} & R_{12} & R_{13} & x_{1,1} \\ R_{21} & R_{22} & R_{23} & x_{1,2} \\ R_{31} & R_{32} & R_{33} & x_{1,3} \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{3.31}$$

Similarly, we create $\mathbf{M}_2$ using $\underline{q}_{n+1}$ and $\mathbf{x}_{n+1}$. The transformation (a combined rotation and translation) that maps $\mathbf{M}_1$ to $\mathbf{M}_2$ is the matrix $\mathbf{A}$ given by

$$\mathbf{A} = \mathbf{M}_2 \mathbf{M}_1^{-1}. \tag{3.32}$$

All coordinates in the reference nucleosome are then transformed by $\mathbf{A}$, the new coordinates are given by

$$\begin{pmatrix} x'_{i,1} \\ x'_{i,2} \\ x'_{i,3} \\ 1 \end{pmatrix} = \mathbf{A} \begin{pmatrix} x_{i,1} \\ x_{i,2} \\ x_{i,3} \\ 1 \end{pmatrix}, \tag{3.33}$$

where i is the index of the i-th bead. The new DNA quaternions are created by transforming their matrix representation by the rotational component of $\mathbf{A}$

$$\mathbf{R}' = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \mathbf{R}. \tag{3.34}$$

The R's are then changed to the quaternion representation giving $\underline{q}'_i$ and the new coordinates and quaternions are appended to chromatin structure. If another nucleosome is added, this procedure is repeated with the new values of $\mathbf{x}_n$ and $\underline{q}_n$.

In practice, to avoid steric clashes, we first pull the linker DNA of the reference nucleosome such that the created chromatin will be in a "beads-on-a-string" configuration; this is illustrated in figure 3.8.

3.7 Computational implementation

The model we have described so far is in a general form and completely described by the potential energy functions. We now need the ability to sample the equilibrium configurations of the system in the Canonical ensemble. The two main methods for achieving this are Monte Carlo and molecular dynamics (with an appropriate thermostat).

The Rigid base-pair model is typically used in MC simulations — key reasons for this are that the potential energy is a simple calculation, whereas, the forces are less simple, namely the torque needs to be computed numerically rather than analytically. Additionally MC sampling enables highly effective sampling of long (but low density) polymer systems by use of pivot moves, or enhanced sampling trial moves such as recoil growth [98]. Our system does involve long polymers, for which MC would be ideal, however we expect the system to reach high densities, which mean MC moves such as pivot moves, will be very hard to achieve with non-zero acceptance probability, e.g. any global MC move will result in high energy molecule overlaps. In our previous unpublished work we did begin using MC methods, and we found that at physiological salt the densities become too large for this to be a sensible sampling choice as the only MC moves that could be accepted were local positional moves e.g. the displacement of a single base-pair. The global moves, that make MC worthwhile, could not be successfully implemented. This lead us to consider a Molecular Dynamics (MD) framework.

MD enables simulations at higher density because during a timestep all atoms move following thermal motion and the potential gradient. Additionally, there are a wealth of established MD codes we can use whilst for MC there are far less established, generalized codes as practitioners usually implement bespoke codes. Using MD means we do not have to implement the complex techniques that allow for good scaling and high levels of parallelism, e.g. cell list domain decomposition. The non standard parts of our model are the DNA bonds, the KH potential, the use of ellipsoidal particles, and the use of rigid bodies. Existing MD codes have varying levels of complexity in implementing new potentials and varying support for including finite size particles and rigid bodies. LAMMPS [94] is a code which already includes ellipsoids and rigid bodies, furthermore its design allows for easy inclusion of new potentials, therefore we have implemented the model in LAMMPS. An important point is that the model could be implemented in other existing MD codes, with varying degrees of difficulty.

3.7.1 Pairwise potential cutoff terms

The pairwise terms we listed in section 3.4.2, $E_{\text{Electrostatic}}$ and E_{KH}, are short range potentials, that is, they converge to zero within distances that are small with respect to the size of the entire simulation box. This feature is essential for the efficient implementation of molecular dynamics in LAMMPS — to avoid the N^2 operation of summing over all pairs of particles the potentials can be 'cutoff' (set to zero) for pairs which have a separation larger than the chosen cutoff distance. LAMMPS uses neighbour lists and the cell list method to achieve this. We will not discuss the full details here, see [94], but simplistically each atom stores a list of which other atoms are within the cutoff distance of itself. The lists do not needed to be updated every

timestep if the list cutoff is greater than the potential cutoff. To update the list each atom does not need to search over the entire simulation box, it only needs to search within a certain domain, or cell, within the simulation box. This is the cell list method: the simulation box is subdivided into cells, which have edge lengths equal to or greater than the cutoff distance, and each atom belongs to one cell. To find its neighbours an atom only has to search in adjacent cells. The has the effect of reducing the order N^2 scaling of molecular dynamics into an order N problem, provided the particle density is homogeneous.

This means the forms of the potentials actually used in LAMMPS are

$$E^{\text{LAMMPS}}_{\text{Electrostatic}} = \begin{cases} E_{\text{Electrostatic}}, & r \leq r_c, \\ 0, & r > r_c, \end{cases} \tag{3.35}$$

$$E^{\text{LAMMPS}}_{\text{KH}} = \begin{cases} E_{\text{KH}}, & r \leq r_c, \\ 0, & r > r_c, \end{cases} \tag{3.36}$$

where r_c is the cutoff distance. For the electrostatic term we use a cutoff distance of 3.5 times the Debye length [137] and for the KH term we use a cutoff distance of 3 times σ, as done in [133].

3.7.2 Implementation in LAMMPS

LAMMPS simulations are typically created using two files: the data file which contains the initial structure of the simulation, namely coordinates of the atoms, bond topology, and in our case the orientation and size of the ellipsoidal particles; and the input script which contains the information about the potentials, the integration method, and how long the simulation should run for. The full details on how to use LAMMPS can be found in its documentation. In this section we describe and explain the typical data files and input scripts used for our model, therefore by necessity this section is aimed at readers who have familiarly with LAMMPS. In this section we will write LAMMPS input script commands in the following form:

```
Example command for a LAMMPS input script
```

The general structure of our LAMMPS setup is as follows. We use real units

```
units real
```

this means:

- distance = Angstrom
- time = femtosecond
- mass = grams/mole
- energy = kcal/mol
- charge = multiples of electron charge (electron = -1, proton = 1)
- temperature = Kelvin

The atom style is

property	data type
atom type	integer
position	3d vector
quaternion orientation	unit quaternion (4d vector)
molecule ID	integer
charge	floating point number
mass	floating point number

Table 3.5: LAMMPS atom properties

```
atom_style hybrid ellipsoid angle charge
```

this means atoms have the properties listed in table 3.5. An important point is that atoms can be denoted as not being ellipsoids, this means they do not have orientation or shape and are treated as standard point particles.

For bond style we use a hybrid style of the standard harmonic bond, our custom DNA bond type, and the zero bonds type.

```
bond_style hybrid harmonic harmonic/DNA zero
```

`harmonic` is the standard LAMMPS harmonic bond, equation 3.14, we use it for all protein bonds. The histone tail bonds are all bond type 1 and have the parameters we mention in section 3.3.1:

```
bond_coeff 1 10 3.8
```

The elastic network model bonds also use this bond style, they each have a unique bond type with number $n > 3$ and the bond coefficients are

```
bond_coeff n 10.0 r
```

where r is the equilibrium bond length of that specific bond in the elastic network model. `harmonic/DNA` is our implementation of the rigid-base pair potential. The source code files `bond_harmonic_DNA.cpp` and `bond_harmonic_DNA.h` are available from our code repository (see section 1.5). This bond style reads in two other files `NAFlex_params.txt` which contains the mean values and stiffness matrices for each of the 16 base-pair steps and `DNA_sequence.txt` which contains the list of which base-pair each DNA bead represents. The bond coefficients are set to zero,

```
bond_coeff 2 harmonic/DNA 0 0
```

they have no effect and are simply left over from the implementation of the standard harmonic bond type that we modified for our implementation. `zero` is a bond style where there is no potential or forces and only the topology is recorded. We use this to ensure that DNA beads do not have pairwise interactions with the DNA beads they are bonded to, as explained in figure 3.3. Specifically the phosphate beads on base-pair i are each bonded, using bond style zero, to both of the phosphates on base-pair $i + 1$ and $i - 1$. Used in combination with the command

```
special_bonds fene
```

which turns off pair-wise interactions between beads that are bonded together.

Moving on to the pairwise interactions we use a custom pair style which includes both the screened Coulomb interaction (equation 3.11) and the KH interaction (equation 3.16). This custom potential is a modified form of `pair_style`

`lj/cut/coul/debye` which contains a screened Coulomb term and a LJ term, the modified version, which was first created by Dignon et al [133] and shared with us, incorporates the lambda parameter into the LJ term and is called `pair_style ljlambda` and is implemented in the files `pair_ljlambda.cpp` and `pair_ljlambda.h` available from our code repository. In the LAMMPS script it is used as

```
pair_style ljlambda k LJCUT COULCUT
```

where k is the inverse debye length, LJCUT is the cutoff distance of the KH interaction and COULCUT is the cutoff distance for the Coulombic term. The pair style coefficients are listed as

```
pair_coeff i j ϵ σ λ LJCUT COULCUT
```

where i and j are the atom types; ϵ, σ, and λ are same as equation 3.16; LJCUT and COULCUT are the cufoff distance for the specific $i-j$ pair. We note here that the atom types run from 1 to 42. 1 to 20 are for the protein beads that are histone tails, each type corresponds to one of the 20 amino acids. 21 to 40 are for the amino acids that are part of globular domains, where once again each type corresponds to one of the 20 amino acids. 41 is the DNA ellipsoids and 42 is the DNA phosphates. The full details are in table D.1.

Supplementing the pair style, the dielectric constant is set to 80 to represent water.

```
dielectric 80
```

The neighbour list skin distance is set to 10 Å,

```
neighbour 10 bin
```

this skin distance is used for building the neighbour lists, all atoms pairs within a cutoff distance equal to the potential cutoff distance plus the skin distance are stored in the list. When any atom moves further than half the skin distance the neighbour lists are rebuilt. Intra-molecule interactions are turned off for globular domains and DNA beads (the composite of an ellipsoid and 2 point particles). For implementation reasons each DNA bead is classed as a different molecule, that is they have unique molecule id labels.

```
neigh_modify exclude molecule/intra globular_domain
neigh_modify exclude molecule/intra dna
```

We use a tiled communication style and dynamic load balancing using the recursive coordinate bisectioning (RCB) method

```
comm_style tiled
fix bl all balance 1000 1.0 rcb
```

This means every 1000 timesteps the domain decomposition (which atoms belong to which processors) is optimized to reduce imbalance in the amount of computation done by each processor. For our systems which have heterogeneous density with respect to the orthogonal simulation box this is crucial for computational efficiency.

The integration settings are different for the protein particles and the DNA particles. For the proteins we use the standard NVE integrator (this uses a velocity-verlet scheme) combined with the langevin thermostat with a temperature of 300 K and a damping time of 10,0000 fs.

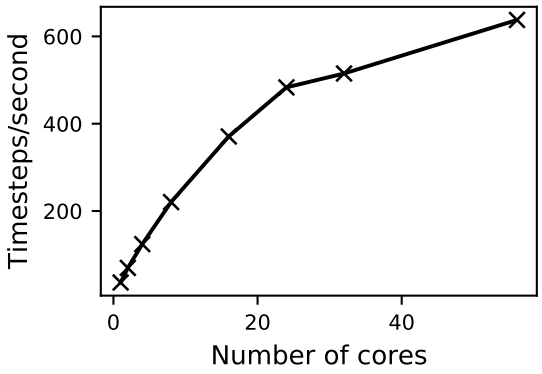

Figure 3.9: Performance of the chemically-specific model for 12-nucleosome chromatin simulations in timesteps per second.

```
fix 1 protein nve
fix 2 protein langevin 300.0 300.0 100000.0 SEED gjf yes
```

The variable `SEED` is the seed for the random number generator, `gjf yes` means we use the GJF formalism [99] which enables the use of a slightly larger timestep. Note we use an older version of fix Langevin before the gjf setting was modified to the version in the current LAMMPS master branch, this is included in our provided code (see 1.5). For the DNA we use a rigid body integrator and a Langevin thermostat,

```
fix 3 dna rigid/nve/small molecule
fix 4 dna_ellipsoids langevin 300.0 300.0 100000.0 SEED angmom 3.0
```

The molecule keyword means each molecule (DNA bead) is treated as a separate rigid body. The Langevin thermostat is applied to the DNA ellipsoids, the `angmom` keyword means that the rotational degrees of freedom are included in the thermostating procedure. We typically use a timestep of 40 fs which we determined is the maximum possible value that ensures simulation stability.

```
timestep 40.0
run N
```

where N is the total number of timesteps to run a simulation for. Finally for some of our simulations we use periodic boundary conditions and for others we use the shrink wrap boundary conditions (essentially there are no boundaries, or it can be thought of as a NVT simulation where the volume is significantly larger than the molecule size). We use the former when there are multiple separate molecules in the system and the latter when we are simulating a single polymer (e.g. a single chromatin fiber) that has no chance of separating and diffusing away as separate components. Complete input scripts for our simulations are available in our code repository.

The performance of a 12-nucleosome chromatin simulation is shown in figure 3.9 as a function of the number of processor cores.

3.8 Enhanced sampling — hamiltonian replica exchange

Preliminary simulations of the model indicated that at physiological salt conditions of 0.15 mol/L NaCl, chromatin condenses into dense structures, additionally the chromatin system is a highly branched polymer, consequently the free energy landscape is populated by many competing low-lying minima separated by high energy barriers. Such a rugged energy landscape is difficult to sample with standard MD simulations, as transitions across the high energy barriers are rare within the accessible simulation timescales. To overcome this we developed a Debye-length Hamiltonian replica exchange method (HREMD) which varies the Debye length, and hence salt screening, across replicas. The weakly screened (larger Debye length) chromatin replicas have extended de-condensed configurations (due to the DNA-DNA electrostatic repulsion), which are easy to fully sample with MD. This is directly analogous to the high temperature replicas of standard TREMD. The reason why we use this HREMD over standard TREMD is twofold. Firstly, the number of replicas needed is lower: a single nucleosome has approximately 1000 protein beads and, assuming a NRL of 180bp, has 180 DNA base-pairs. Each protein has 3 degrees of freedom, each DNA ellipsoid has 6 (3 displacements plus 3 rotational). This results in 4080 degrees of freedom per nucleosome, for a 12-nucleosome chromatin system this becomes 48960 degrees of freedom. Putting this into a TREMD temperature replica generator [116] reveals that 80 replicas are needed to span the temperature range 300K to 600K with an acceptance probability of 0.3. For our HREMD method we find that 16 replicas are sufficient to cover a large enough chromatin compaction range with acceptance probabilities of 0.3. Secondly, all the replicas used in our HREMD simulations are at physically meaningfully salt conditions and at 300K, this means all replicas can be used for analysis, whilst in TREMD only the 300K replica can be used (with additional reweighing procedures [138] other replicas could be used with decreasing significance as the temperature increases).

The exchange probability between replica i and $i+1$ for the HREMD method is given by:

$$\begin{aligned} P(i \leftrightarrow i+1) &= \min\left(1,\, \exp\left[\frac{1}{k_B T}\Delta\right]\right), \\ \Delta &= \left(U_{\lambda_D^i}(\mathbf{x}_i) - U_{\lambda_D^i}(\mathbf{x}_{i+1}) + U_{\lambda_D^{i+1}}(\mathbf{x}_{i+1}) - U_{\lambda_D^{i+1}}(\mathbf{x}_i)\right), \end{aligned} \tag{3.37}$$

where $\mathbf{x}_i$ are the chromatin coordinates of the i^{th} replica, and $U_{\lambda_D^i}$ the potential energy function with Debye length λ_D^i. The exchange is accepted or rejected based on the Metropolis criteria, and upon exchange the potential energy functions are switched. For our 12-nucleosome chromatin systems, which we present in the next chapter, we find that at the value of $\lambda_D = 8.0\,\text{Å}$ (corresponding to 0.15 mol/L salt), the chromatin structures are compact and suffer from sampling issues, increasing the Debye length to 15 Å gives rise to open structures that can sample effectively. We use 16 replicas in the range 8.0–15.0 Å giving an acceptance probability of 0.3.

The difference between $U_{\lambda_D^i}$ and $U_{\lambda_D^{i+1}}$ lies only in the $E_{\text{Electrostatic}}$ term, thus at each exchange attempt step we first need to compute the current value of $E_{\text{Electrostatic}}$ using λ_D^i and then we compute the value of $E_{\text{Electrostatic}}$ again using the value of λ_D^{i+1}.

We implement our HREMD method by modifying an existing version of LAMMPS

parallel tempering command from the `REPLICA` package. The source files `hremd.cpp` and `hremd.h` are available in our code repository.

Chapter 4

Simulations of chromatin at DNA base-pair and amino-acid resolution

In this chapter, we report results from our chromatin simulations using the chemically-specific model described in the previous chpater. We present a combination of validation/testing exercises and novel results. Part of these results are in [121].

Contents

4.1 Determination of the simulation timesteps

To determine the appropriate timestep to use in our simulations, we first looked at systems of just DNA because this includes our newly implemented rigid base-pair potential. Thus, the appropriate timestep for this model is unknown. Subsequently, we look at the behavior of our full chromatin model with simulations of a single nucleosome.

4.1.1 DNA model timestep

The stiffest parts of the DNA potential come from the rigid base-pair potential so this will determine the maximum timestep we can use. Looking at the stiffness

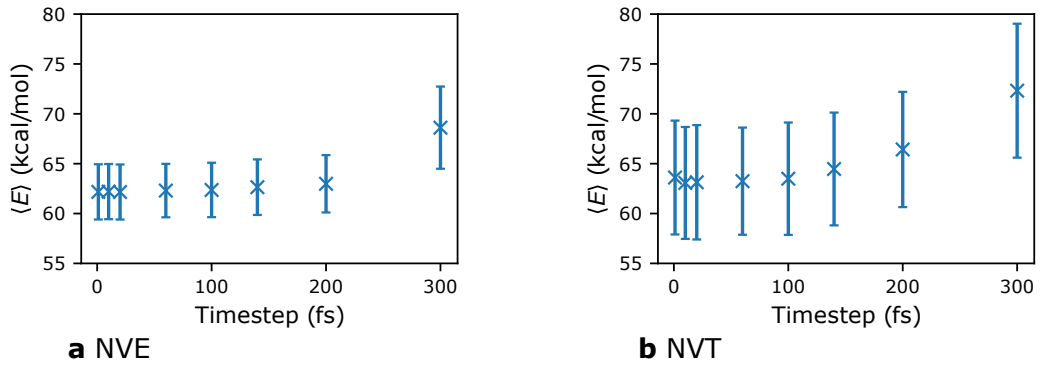

Figure 4.1: **Determination of timestep for the DNA model. (a)** Average total energy for NVE simulations using different timesteps. **(b)** Average total energy for NVT simulations using different timesteps. The data points in both subplots are the mean±standard deviation.

matrices the largest displacement force constants are approximately 10 kcal/mol/Å^2. The period corresponding to this harmonic term is

$$t = 2\pi\sqrt{m/k}, \tag{4.1}$$

which for our DNA beads mass of 650 g/mol gives $t = 2.5$ ps. Checking the angular terms we see that they are approximately 0.03 kcal/mol/degree^2, with a corresponding angular frequency,

$$\omega = \sqrt{k/I}, \tag{4.2}$$

where I is the moment of inertia. For our ellipsoid DNA beads, the smallest component of I is about the y axis and is given by $I = \frac{1}{5}m(a^2 + c^2)$ where a and c are 5.5 Å and 1.75 Å respectively. This gives an approximate period of 2 ps. These two timescales suggest 200 fs as a sensible upper limit on the timestep, i.e. an order of magnitude smaller. To check this, we ran NVE simulations of a single 100 bp coarse-grained DNA strand to measure energy conservation as a function of timestep. All simulations were initialized in identical states at 300 K and then run for 1000 ns using the respective chosen timestep. The mean total energies (potential + kinetic) are plotted in figure 4.1a. We see that, as expected, there is significant divergence of the mean energy for a timestep greater than 200 fs, while for timesteps less than 200 fs, the energies are indistinguishable. We repeated the same procedure in the NVT ensemble using a Langevin thermostat with a damping period of 100 ps, the mean energies are plotted in figure 4.1b. We see that a timestep of 100 fs gives the same energies as a timestep of 1 fs. Thus we choose 100 fs as the appropriate timestep for the DNA model.

4.1.2 Chromatin model timestep

For the protein model, from the work of Dignon et al [133], we know that a suitable time-step is 10 fs. To verify this, we ran a NVE simulation of a single nucleosome and verified that the energy is conserved; this is plotted in figure 4.2a. It is often the case that when using Langevin dynamics (which we will be doing for all production simulations), one can moderately increase the timestep; this is due to the damping

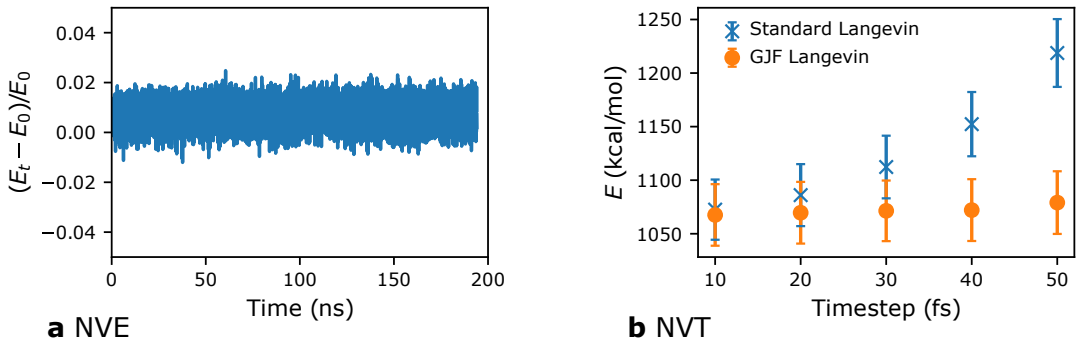

Figure 4.2: **Determination of timestep for the chromatin model. (a)** Energy time-series for a NVE simulation of a nucleosome demonstrating energy conservation for a timestep of 10fs **(b)** Average energy from NVT simulations of a nucleosome, as a function of timetep, using two different Langevin thermostat settings. The data points are the mean±standard deviation of the potential energy.

term reducing large unstable velocities. To investigate this, we ran the single nucleosome system in the NVT ensemble using a Langevin thermostat with a damping time of 100 ps and measured the average potential energy as a function of timestep, which is plotted in figure 4.2b. We found that the standard Langevin thermostat in LAMMPS induces a noticeable change in the energy when the timestep is increased. However, when using the GJF [99] formulation, the measured energy for a timestep of 40 fs is indistinguishable from the energy with a timestep of 10 fs. This lead us to choose a timestep of 40 fs, provided we use the GJF formulation, for our chromatin model simulations.

4.2 Estimation of DNA persistence length

To test our chemically-specific coarse-grained DNA model, we computed the persistence length of DNA as a function of monovalent salt concentration (Figure 4.3a) and DNA sequence (Figure 4.3b). The simulations were performed using 300 bp long strands of isolated DNA. Ten independent simulations were performed for each data point, for a total simulation time of 100 million time steps per system.

The persistence length P of a polymer is described as the length at which correlations in the direction of the polymer tangent are lost. We use the following definition:

$$\begin{aligned} \langle \mathbf{n}(x_i) \cdot \mathbf{n}(x_j) \rangle &= e^{-l/P}, \\ l &= |x_i - x_j|. \end{aligned} \tag{4.3}$$

Here, $\mathbf{n}(x_i)$ is the tangent vector of the polymer at position x_i and l is the contour distance in units of base-pair (bp). Note that the position is given in terms of the polymer contour distance with units of base-pairs (symbol bp), e.g., $x_1 = 1$ is the 1st base-pair and $x_{10} = 10$ is the 10th base-pair. To compute the value of P from our simulations, we use the DNA ellipsoids quaternion orientation to directly give the tangent vectors. The 20 base-pairs at each end of the strand were excluded from the analysis to eliminate any edge effects. For each timestep, all pairs of $\mathbf{n}(x_i) \cdot \mathbf{n}(x_j)$ were computed and the average was taken over all timesteps for each l value. We

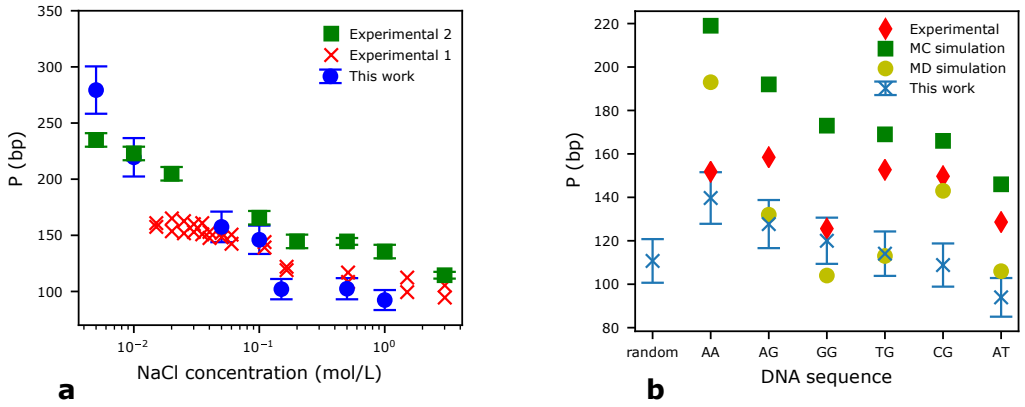

Figure 4.3: **Comparison of DNA model persistence length with experimental values. (a)** Persistence length of DNA in units of base-pairs (bp) as a function of NaCl concentration in mol/L at 300K. Blue circles are the values obtained with our simulations on a set of 300 bp DNA strands. Red crosses (experimental 1) are values from single-molecule high-throughput tethered particle motion experiments on DNAs of 1201 and 2060 bp at room temperature [139]. Green squares (experimental 2) are values from Rayleigh light scattering experiments for a T7 bacteriophage DNA from [140]. **(b)** Persistence length of DNA as function of DNA sequence from our simulations (blue crosses) using a random sequence and six poly(XY) sequences for DNA of 300 bp in length. We compare to values from experimental cyclization assays [141], coarse-grain Monte Carlo simulations [142], and all-atom MD simulations [142]. The data points are the values of P fitted by a non-linear least squares fit of equation 4.3, the error bars are the standard error in P reported by the least squares fitting method. 10 independent simulation trajectories were used.

then plot the scatter graph of l versus $\langle \mathbf{n}(x_i) \cdot \mathbf{n}(x_j) \rangle$ and fit an exponential curve to obtain P.

In figure 4.3a we have compared the salt dependent DNA persistence length measured with our model with two different sets of experimental values available in literature. Experiment 1 [139] are values computed from single molecule high-throughput tethered particle motion experiments on DNA strands of 1201 and 2060 bp at room temperature. Experiment 2 [140] are values computed from Rayleigh light scattering experiments for a T7 bacteriophage DNA. We see reasonable agreement with our values and those from literature, the differences between our values and the experimental ones are of the same order as the differences between those of the two experimental plots. This highlights the difficulty in accurate measurement of the persistence length of semi-flexible polymers with the existence of multiple techniques and alternative definitions to our Eq. 4.3 [143–145].

In figure 4.3b we have plotted the sequence-dependent behavior of the persistence length, our results follow the general trend where stiffness decreases going in the order: [poly(AA), poly(AG), poly(GG), poly(TG), poly(CG), poly(AT)]. All but one data point are in agreement with either the experimental and/or MD results. The experimental results [141] are values computed from cyclization assays. The MD simulation values [142] are from all-atom MD simulations of DNA of length 10-20 bp. The MC simulation values [142] are from Monte Carlo simulations using a rigid-base coarse-grained model of DNA for length 220 bp. We note that there are significant differences between all the different reported sequence-dependent values of DNA persistence lengths stemming from experiments and simulations, which is consistent

with the aforementioned difficulty in accurate persistence length calculations.

4.3 Testing the HREMD method on a small DNA circle

We first tested our HREMD method on a small 200 bp supercoiled DNA circle (over-twisted by 10%), using our DNA coarse-grained model. We chose this system because it can be sufficiently sampled across a wide-range of salt concentrations, without the need of using enhanced sampling techniques. This system allowed us to verify that our HREMD implementation was correct, i.e that it satisfies detailed balance and the replicas correctly sample the distributions corresponding to the Hamiltonian we give them. As required, we observe that at high salt, the supercoiling dominates and, the small supercoiled DNA circle adopts a figure-of-eight shape (plectonemic supercoiling) [66, 146, 147]. At low salt, in contrast and also as expected [146, 148], the DNA–DNA repulsion dominates and the system adopts a circular shape. For these tests, we ran the HREMD method using 16 replicas spanning a Debye length from 8 to 30 Å. We chose the replica schedule using a geometric sequence, shown in table 4.1, which produced exchange probabilities of 0.4.

Replica	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
λ_D (Å)	8.00	8.74	9.54	10.42	11.38	12.43	13.57	14.82	16.19	17.68	19.31	21.09	23.03	25.15	27.47	30.00

Table 4.1: Debye-length replica scheme. We have used a geometric sequence of increasing Debye-length.

Figure 4.4 shows the results from the simulation. The potential energy for each Debye length replica is plotted in figure 4.4 a. We see that the end points of the replica range, 8.0Å and 30Å, are indistinguishable from the standard MD simulations carried out at those Debye lengths, confirming that our method correctly obeys detailed balance. In figure 4.4b, we show the radius of gyration for each Debye-length replica, where the existence of two different states — corresponding to the circular and figure-of-eight configurations — is clear. In figure 4.4b, we have plotted the time series of each replica's Debye-length values, demonstrating that the simulation trajectories effectively travel through replica space.

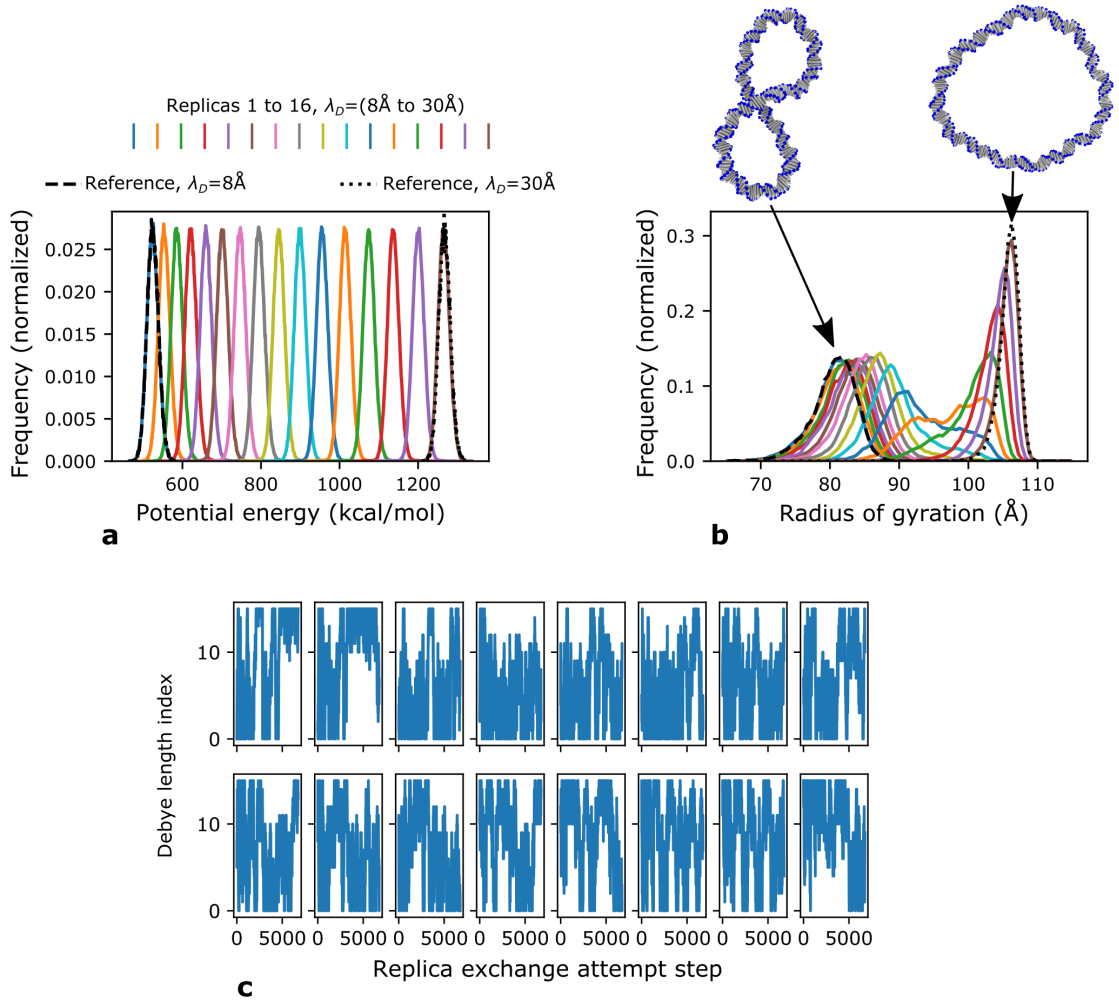

Figure 4.4: **Testing the HREMD method on the salt dependence of DNA mini-circle supercoiling**. **(a)** The potential energy distributions for each λ_D, computed from using the HREMD method, with overlayed curves from standard MD simulations at the λ_D end points (labelled reference). We see that the distributions calculated from HREMD and standard MD are in agreement. **(b)** Radius of gyration distributions for each λ_D, the ordering of the peaks from left to right follows the same order as subfigure a, which is the order of increasing λ_D. The supercoiled figure of eight state and the circle state are clearly distinguishable as the two populations and are labelled. **(c)** Time series of the λ_D index for each replica, showing good traversal of replica space.

4.4 Nucleosome formation

In this section, we test the capacity of our chemically-specific model to capture the spontaneous re-wrapping of nucleosomes (from free DNA and a histone protein octamer) derived from its ability to account for the electrostatic and hydrophobic DNA-protein interactions, and the mechanical properties of the DNA.

4.4.1 Completely unrestrained DNA

We first investigated the ability of the model to spontaneously form nucleosomes from an initial state where the DNA is completely unbound from the histone core, but while forcing the histone protein octamer to remain fully formed. The simulations were performed by setting up a 211 bp DNA strand in a linear configuration at a distance of more than 150 Å from a randomly rotated histone core, as pictured in figure 4.5a. A periodic simulation box was used to ensure the DNA and protein eventually come into close enough contact for binding to occur. The box dimensions were set to 800 Å cubed, which is greater than the length of the DNA strand by approximately 100 Å; this ensures that the molecule will never see the periodic copy of itself. We performed 64 repeats of this simulation, each with different random starting orientations, at a salt concentration of 0.15 mol/L NaCl and a temperature of 300K. The simulations were run for 100,000,000 timesteps (4µs) with coordinate snapshots recorded every 10,000 timesteps, and energies recorded every 1000 timesteps.

An example time series of the potential energy for one of the simulations is shown in figure 4.6a. The sharp drop in the energy indicates the moment in time when a binding event occurs. We computed the time taken for DNA binding to take place, or t_{bind}, in all the 64 simulations (figure 4.6b), and found that for 46 simulations out of the 64 $t_{\text{bind}} < 0.2\,\mu\text{s}$, for 63 simulations out of 64 $t_{\text{bind}} < 1.0\,\mu\text{s}$, and for one simulation $t_{\text{bind}} = 1.6\,\mu\text{s}$. This indicates that the binding events are well within the reach of timescales that can be comfortably investigated with our model.

An important observation is that the the final DNA-histone core bound configurations obtained across the different simulations are heterogeneous. These structures were categorised by visual inspection into the categories shown in the pie chart in figure 4.5a. We observed that a canonical nucleosome (i.e., with left-handed DNA supercoiling) is only correctly formed in approximately one third of the attempts, the other resulting configurations include: reversomes —i.e., where the DNA is wrapped in a right handed helix instead, "knotted" configurations — where the DNA coiling is nucleosome-like but has DNA overlap, and "wrong" configurations — where the DNA is bound to the histone but there is no nucleosome-like coiling. This suggests that additional processes or constraints are required to ensure the model always forms nucleosomes. This result is not surprising if one considers that the assembly of nucleosomes in vivo relies on the action of histone chaperones and other enzymes [149, 150], and in vitro [151, 152], on the use of slow salt gradients and cycles of heating and cooling down; the latter which are aimed at relaxing metastable DNA-histone configurations into the canonical nucleosomes, which represent the global minimum. Experimental literature shows that the formation of nucleosomes is significantly slowed when completely relaxed DNA is used. Futhermore, when positively supercoiled DNA nucleosomes is used, nucleosomes fail to assemble, but

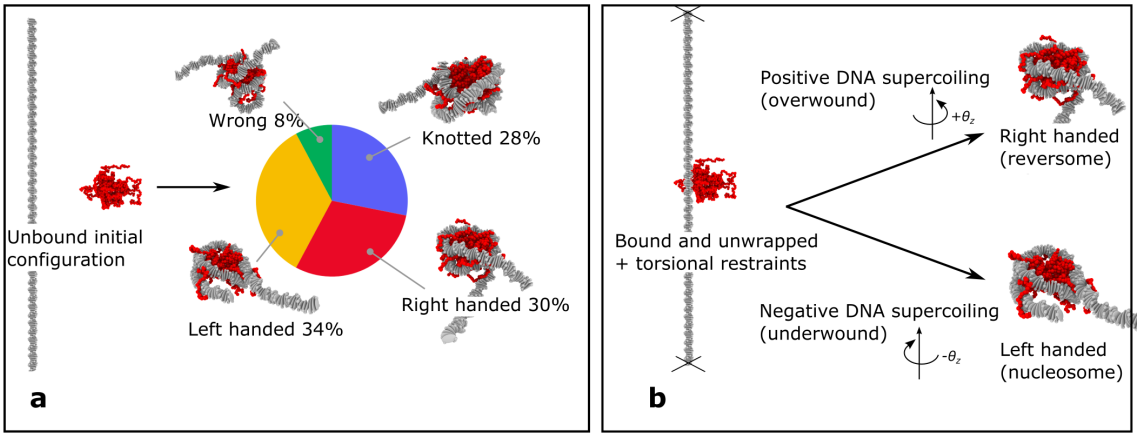

Figure 4.5: **Nucleosome formation. (a)** Starting from an unbound configuration, as time progresses, the DNA wraps around the histone core in one of four possible ways: left-handed supercoiling (a canonical nucleosome) with a 34% probability, right-handed supercoiling (a reversome) with a 30% probability, a knotted configuration which is neither a left-handed or right-handed supercoil but still resembles a nucleosome-like shape with a 28% probability, and other configurations which differ significantly from nucleosome topology ("wrong") with a 8% probability. **(b)** Starting with a fully unwrapped nucleosome, but with the histone core bound to the DNA at a single point, we apply a weak super-coiling and torsional restaraints. In this case, as time progresses, the model always form nucleosomes when DNA is underwound, and reversomes when DNA is overwound.

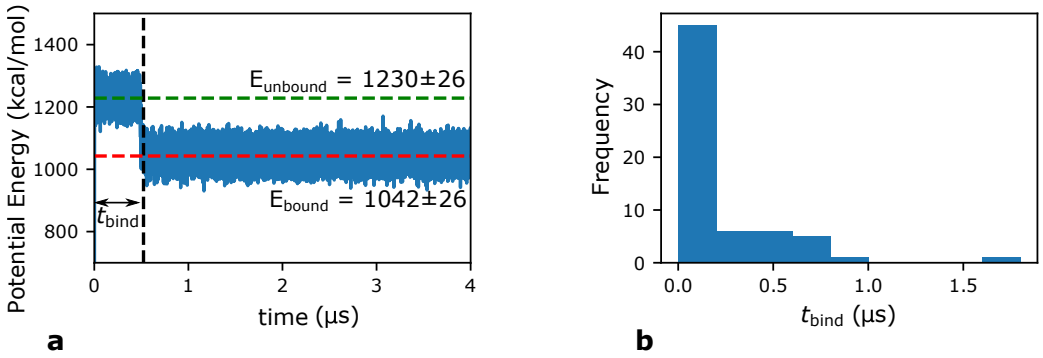

Figure 4.6: **Nucleosome formation binding time. (a)** Time series of potential energy for one of the nucleosome formation simulations. The drop in potential that occurs at t_{bind} is when the DNA binds to the protein. The mean potential energy of the unbound state, green line, is 1230 ± 26 kcal/mol; the mean potential energy of the bound state, red line, is 1042 ± 26 kcal/mol. The uncertainties are the standard deviations. **(b)** Histogram of t_{bind} for the 64 simulations.

when negatively supercoiled DNA is present instead, the nucleosomes form instantaneously [153]. To investigate the response of the model to DNA supercoiling, we repeated our assembly simulations with the addition of torsional restraints to the DNA, which was setup to give rise to either negative or positive supercoiling. The results of these simulations are discussed below.

4.4.2 Nucleosome formation with torsional restraints

The DNA supercoiling was achieved by either increasing (for positive supercoiling) or decreasing (for negative supercoiling) the twist angle between successive base-pairs by 10%. The torsional restraints fix the rotation of the last two base-pairs by the addition of strong forces, such that these base-pairs cannot rotate about their respective z-axis. Additionally, we started the simulations with the DNA in contact with the histone core. We performed 64 repeats of these simulations for each of the two DNA supercoiling configurations, and observed that negative supercoiling always forms a nucleosome and positive supercoiling always forms the chirally inverted reversome (Figure 4.5b), consistent with experimental observations. The reversome is a metastable state of a nucleosome that is observed in experiments [154–156], and has been hypothesized to be present in centromeric chromatin (that is a region of chromatin with unusual nucleosome structures that is involved in the segregation of the chromosome poles) [156].

4.4.2.1 Details of torsional restraints on the nucleosomal DNA

The torsional restraints on the DNA were implemented by adding extra constant forces with a magnitude of 10 kcal/mol/Å to the phosphate particles belonging to the two end DNA beads (DNA bead 1 and DNA bead N). The direction of the force on phosphate 1 of each DNA bead is the $\hat{\mathbf{y}}$ direction of DNA's initial orientation, and the direction of the force on phosphate 2 is $-\hat{\mathbf{y}}$. The forces balance each other out, i.e there is no overall motion applied to the molecule, and the forces only act to prevent any rotation of the DNA beads around their $\hat{\mathbf{z}}$ axis. Motion in the z direction is unrestrained.

4.5 Free-energy cost of single nucleosome unwrapping

An important aspect of our model is the ability for the nucleosomes to dynamically unwrap and re-wrap, as observed in vivo. The forces and energy barriers involved in nucleosome unwrapping have been quantitatively investigated with force spectroscopy experiments [157–160]. The availability of this data, provides an opportunity to test the performance of our model.

In experiments, force-extension curves are computed by means of force spectroscopy methods [157–160]. This involves pulling the nucleosomal DNA at constant velocity, and measuring the force applied by the clamp. The pulling velocities need to be slow enough such that the system is always in equilibrium, typically speeds of the order of milimeters per seconds are needed (e.g. $\sim$0.1 mm/s was used in the magnetic tweezer experiments of Meng et al [160]). In molecular simulations, such slow speeds are not achievable within the available simulation timescales. For our

model, unwrapping a single nucleosome at a pulling speed of ∼0.1 mm/s would require a wall-time of approximately 100 days to cover the extension range of 0–700 Å that is needed to fully unwrap a 211 bp nucleosome. Using faster speeds does not solve this issue, as the measured force depends on the pulling velocity [161]. To overcome this computational limitation, we instead used umbrella sampling simulations to estimate the Potential of Mean Force (PMF) of nucleosome unwrapping.

4.5.1 Simulation methods

To compute the PMFs of nucleosome unwrapping using umbrella sampling simulations we follow a similar method to that implemented by Lequieu et al. [88]. Here, the collective variable is the DNA extension, which is the distance between the first and last base-pairs. To implement the umbrella sampling procedure in LAMMPS, we use the COLVARS [162] (version 2019-08-05) package. Starting from an equilibrium structure with an extension of 25 Å, initial configurations for the windows were prepared via constant velocity Steered MD (SMD) until the extension was at 750 Å. A spring constant of 0.01 kcal/mol/Å^2 was used with a pulling velocity of 9.0×10^{-6} Å/fs, giving a total pulling time of 100 ns. The extension range was split into 50 equally spaced windows. Each window was run with a fixed harmonic biasing potential at the corresponding extension with a spring constant of 0.025 kcal/mol/Å^2 for 100 ns. These values were chosen by assessing the histogram overlap and checking that the calculated PMFs were the same on longer timescales. The entire procedure was repeated 5 times and the aggregate data was used for computing the PMF via the Weighted Histogram Analysis Method (WHAM) [97]. The same procedure was done for all nucleosome configurations and environments.

4.5.2 Results and Discussion

We computed the unwrapping PMFs for six different nucleosome conditions using the NCP147 (PDB:1KX5 [5]) DNA sequence unless otherwise stated: With H1 linker histone at 0.15M NaCl; without Linker histone at 0.05M, 0.15M, and 0.3M; without linker histone at 0.15M and with a poly-A DNA sequence; and without linker histone with all histone tails removed at 0.15M. The resulting PMFs are plotted in figure 4.7a top panel for the different conditions. Subsequently, we derived the force-extension curves by computing the numerical derivative of the PMF curves, plotted in figure 4.7a lower panel.

Consistent with experiments [160], the nucleosome unwrapping behavior predicted by our model can be separated into three states: state 1 corresponds to a fully wrapped nucleosome, state 2 has the first turn of the DNA unwrapped, and state 3 is fully unwrapped but with the histone core still bound to the DNA. These states are indicated by the dashed lines in figure 4.7a and illustrated with simulation snapshots in 4.7b. We estimated the free energy changes between the states, ΔG_1 and ΔG_2, and the corresponding rupture forces, F_1 and F_2, by reading off the graphs; the resulting values are given in table 4.2.

Our value of ΔG_1 for -LH (that is a nucleosome without linker histone) at 0.15M is 11.5 k_BT, which closely matches the value stemming from force-spectroscopy experiments at similar salt conditions (9-11k_BT table 4.3, Refs [158, 160, 164]). Our values of F_1 are also in reasonable agreement with experiments [158, 160, 164]).

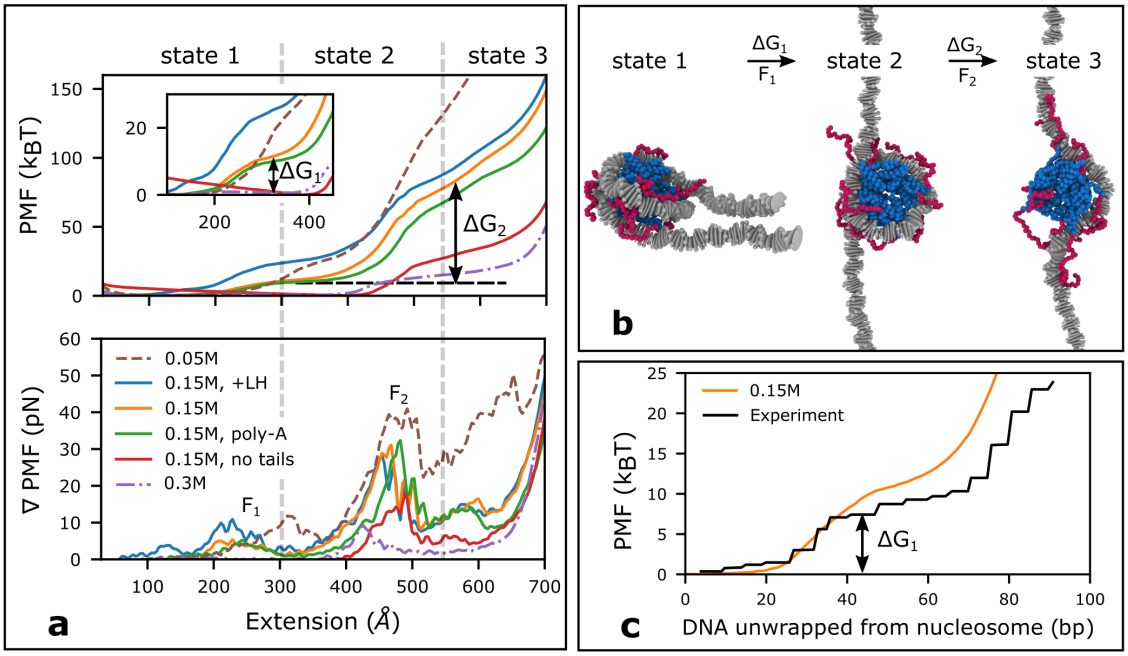

Figure 4.7: **Nucleosome unwrapping**. **(a)** PMFs of nucleosome unwrapping in the top panel with corresponding numerical derivatives plotted in the lower panel. The inset is a zoom in of the low extension regime. The gray dashed vertical lines approximately divide the extension range into the three nucleosome unwrapping states. **(b)** Simulation snapshots showing the three nucleosome unwrapping states. **(c)** Comparison of our computed PMF with the free energy profile from a mechanical DNA unzipping experiment [163]. Here the x-axis has been converted from extension to the amount of DNA that is unwrapped.

Conditions	ΔG_1 (k_BT)	ΔG_2 (k_BT)	F_1 (pN)	F_2 (pN)
0.05M	22 ± 3	133 ± 10	11 ± 1	38 ± 2
0.15M, +LH	25 ± 1	71 ± 5	10 ± 2	28 ± 2
0.15M	11.5 ± 0.6	75 ± 5	5 ± 1	30 ± 2
0.15M, Poly-A	9.9 ± 0.4	66 ± 5	5 ± 1	31 ± 2
0.15M, no tails	-5.0 ± 0.4	30 ± 2	n/a	17 ± 2
0.3M	0.61 ± 0.05	15.8 ± 0.8	1.0 ± 0.5	10 ± 1

Table 4.2: Free energies and rupture forces from our nucleosome unwrapping PMF simulations.

Study	G_1 (k_BT)	F_1 (pN)
Mihardja et al [158]	10	3
Meng et al [160]	9	3-7
Chien et al [164]	11	3

Table 4.3: Values from literature of ΔG_1 and F_1 at similar conditions to our 0.15M simulations.

Force-extension experiments by Spakman et al [165] on 12-nucleosome chromatin reveal that nucleosomes within chromatin fully unwrap at ∼20 pN, this is in agreement with our values of F_2 shown in table 4.2. In figure 4.7c, we directly compare our computed PMF with the free energy profile from mechanical unzipping experiments [163] at single base pair resolution, demonstrating quantitative agreement between our model and the experiment, particularly in the low extension regime.

Additionally, we see an increase in ΔG_1 of $10\,k_BT$ when decreasing the salt concentration from 0.15 M to 0.05 M; this is consistent with the trend from magnetic tweezers experiments [164], which show a change in ΔG_1 of $7\,k_BT$ when decreasing monvalent salt from 0.2 M to 0.01 M. Following the same trend, we see that a high monovalent salt concentration (i.e. 0.3 M), destabilizes the nucleosome as unwrapping of the first turn of DNA can take place with minimal energetic cost. This trend is explained by the electrostatic interactions between the DNA and histone core increasing as the salt screening is reduced.

Further observations are that using a DNA sequence of poly-A, known to be unfavorable to nucleosome formation, instead of the DNA sequence of NCP147, reduces the free energy cost of nucleosome unwrapping by $1.5\,k_BT$ for ΔG_1, and by $9\,k_BT$ for ΔG_2. This is in line with the dependence of nucleosome unwrapping on DNA sequence [166]. We also found that removing the histone tails results in a free energy profile where state 2 is the most energetically favorable, in agreement with experiments where unwrapping of the outer DNA turn occurs at near-zero forces after histone tail removal [157].

Finally, we found that the addition of H1 linker histone increases ΔG_1 by $13.5\,k_BT$ in agreement with force spectroscopy experiments [167].

4.6 Force-extension behavior of chromatin fibers

In this section, we describe the predictions of our model on the force-extension response of 4-nucleosome chromatin arrays. For these systems, umbrella sampling becomes prohibitive due to the large extension range that needs to be covered. Instead, we used a constant velocity steered MD protocol. Because this is a non-equilibrium simulation, we do not expect the forces to be directly comparable to experiments, and are unable to measure free energies. Hence, we discuss our findings qualitatively.

4.6.1 Simulation methods

We used the COLVARS [162] package in LAMMPS to perform the constant velocity steered MD protocol of chromatin fibers. The spring constant was set to $0.001\,\mathrm{kcal/mol/\AA^2}$, and the pulling velocity was 3.0×10^{-6} Å/fs. The initial structures were equilibrated for 100 ns and then the SMD procedure was run for 1000 ns. Values of the applied SMD force and current extension were recorded at each timestep. We repeated this procedure 5 times each for a system containing H1 linker histone (one H1 per nucleosome), and a system with no linker histone. Additionally, we performed one simulation for a system with non-breathing nucleosomes where the nucleosomal DNA is permanently bound to the histone core.

4.6.2 Results and discussion

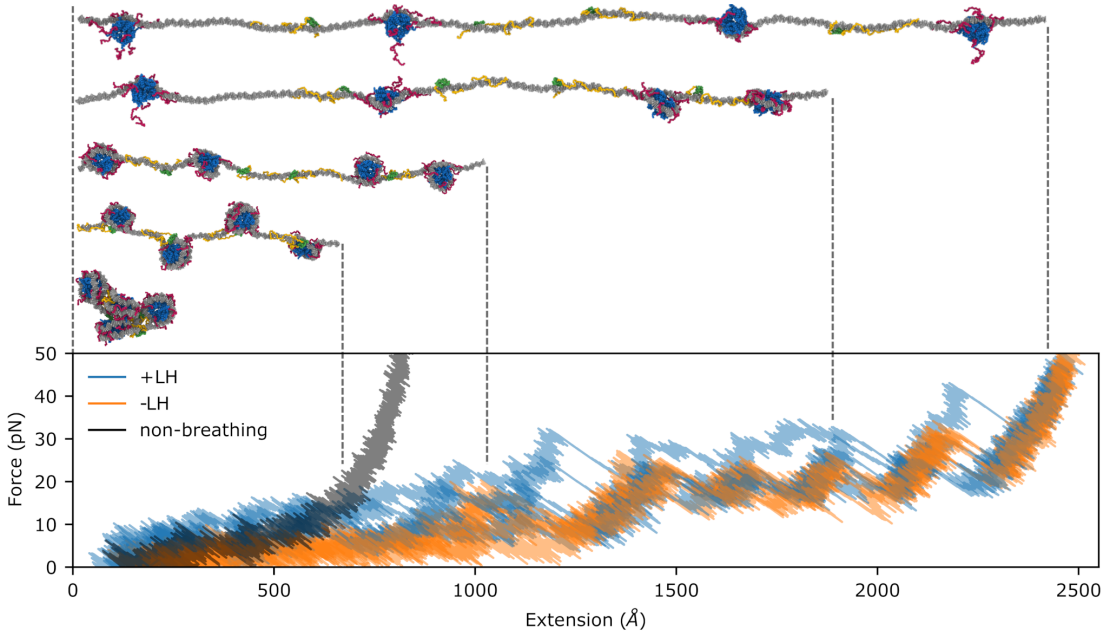

Figure 4.8: **Force-extension of 4-nucleosome chromatin**. In the plot there are 5 curves each for simulations with and without linker histone (+LH and -LH respectively). The non-breathing curve is a simulation where the DNA is permanently bound to the histone core. The simulation snapshots in the upper panel show typical chromatin configurations at the corresponding extension.

In Figure 4.8, we compare the force-extension response of a 4-nucleosome system with and without H1 linker histone (LH). Additionally, we include the resulting curve for the non-breathing model where the DNA is permanently bound to the histone core. After the initial low force regime, where chromatin is pulled into a bead-on-a-string conformation (extension < 600Å), our force-extension curves exhibit the typical saw-toothed pattern observed in optical tweezer experiments of chromatin [168]; that is, the force exhibits peaks followed an abrupt drop accompanied by a certain increase in the extension due to the unwrapping of individual nucleosomes. When we use the non-breathing model, i.e. there is no nucleosome unwrapping by design, and, after the initial low force regime, we purely see the harmonic response of the DNA stretching. Adding linker histones has the effect of increasing the unwrapping forces, both in the low-force regime and the high-force nucleosome unwrapping events.

The force required for unwrapping nucleosomes within chromatin is in the $20-30$ pN range. This agrees with our values of F_2 computed for a single nucleosome in the previous section. Futhermore, this value is in close agreement with optical tweezer force-extension experiments on 12-nucleosome chromatin by Spakman et al [165], who report that nucleosomes within chromatin unwrap at $\sim$20 pN. This agreement implies that our initial concerns about performing the SMD at fast velocities (larger velocities produce larger forces) may be unfounded. One reason for this is that our CG timescales are not equivalent to the "real" experimental timescales, in fact they are much faster, i.e 1 ns in our CG simulations is equivalent to $\gg 1$ ns of all-atom simulation or "real" time.

4.7 Coarse-grained investigation of nucleosome sliding

In the previous sections, we investigated the breathing/unwrapping properties of nucleosomes, now we direct our attention to the phenomenon of nucleosome sliding.

Nucleosome positioning within the genome is significantly controlled by the DNA sequence; this is in part due to the mechanical properties of the DNA, as different sequences exhibit distinct abilities to bend sharply into the nucleosome super-helical conformation. Segal et al [54], proposed that up to 50% of nucleosome positioning in vivo is determined by the DNA sequence. The other important mechanism in play is the dynamic regulation of nucleosome positioning necessary for cell function, which is in part controlled by chromatin remodeler proteins [169].

Interestingly, it has been found that nucleosomes can spontaneously reposition in the absence of any remodeler proteins [170–172]; this suggests that thermal motion can induce nucleosome sliding. There are two main proposed mechanisms to explain how nucleosomes slide across a sequence: 'Twist diffusion' where the DNA undergoes a corkscrew-like motion, moving 1 bp at a time [173, 174], and 'Loop propagation' where DNA loops form on one side of the nucleosome and move around the histone with reptation-style motion [175–177].

To investigate the sequence dependence of nucleosome sliding, we performed single nucleosome simulations for two different DNA sequences: NCP147 [5], a high-affinity nucleosome sequence; and poly-A a low-affinity nucleosome sequence.

4.7.1 Simulation methods

We used a DNA length of 400 bp, by repeating the NCP147 sequence. To eliminate edge effects of the DNA strand and to deal with the case of the nucleosome sliding to the end of the DNA, we bonded the DNA strand to itself across the periodic simulation box boundary. This effectively simulates an infinitely long strand of DNA. The box dimensions are large enough to ensure the nucleosome particle will never interact with itself. The simulation setup is pictured in figure 4.9b. The simulations were run in the NVT ensemble with a Langevin thermostat at 300K. To assess the extent of nucleosome sliding driven by the two different DNA sequences we recorded the position of the nucleosome dyad (in terms of which DNA base-pair is currently as the dyad location) throughout the simulations. The nucleosome dyad is the location of the central nucleosomal DNA base-pair, labelled in figure 1.1b. We defined its location by first computing which DNA base-pairs are in contact with the histone core, and then taking the median base-pair to be the dyad location. We then computed the Mean Squared Displacement MSD of the dyad position as

$$\mathrm{MSD}(t) = \langle |\mathbf{x}(0) - \mathbf{x}(t)|^2 \rangle, \tag{4.4}$$

where $\mathbf{x}(t)$ is the position of the nucleosome dyad at time t in units of DNA base-pairs (this is a 1d vector). We ensured that if the nucleosome moved a full box length relative to $\mathbf{x}(0)$, then $\mathbf{x}(t)$ was computed as the unwrapped displacement, not the displacement wrapped to the periodic boundaries. We performed 32 repeats, each run for 10µs, we then split each of these trajectories into 5 sets of 2µs trajectories giving 160 trajectories to average over. The MSD of a 1D diffusion process can be

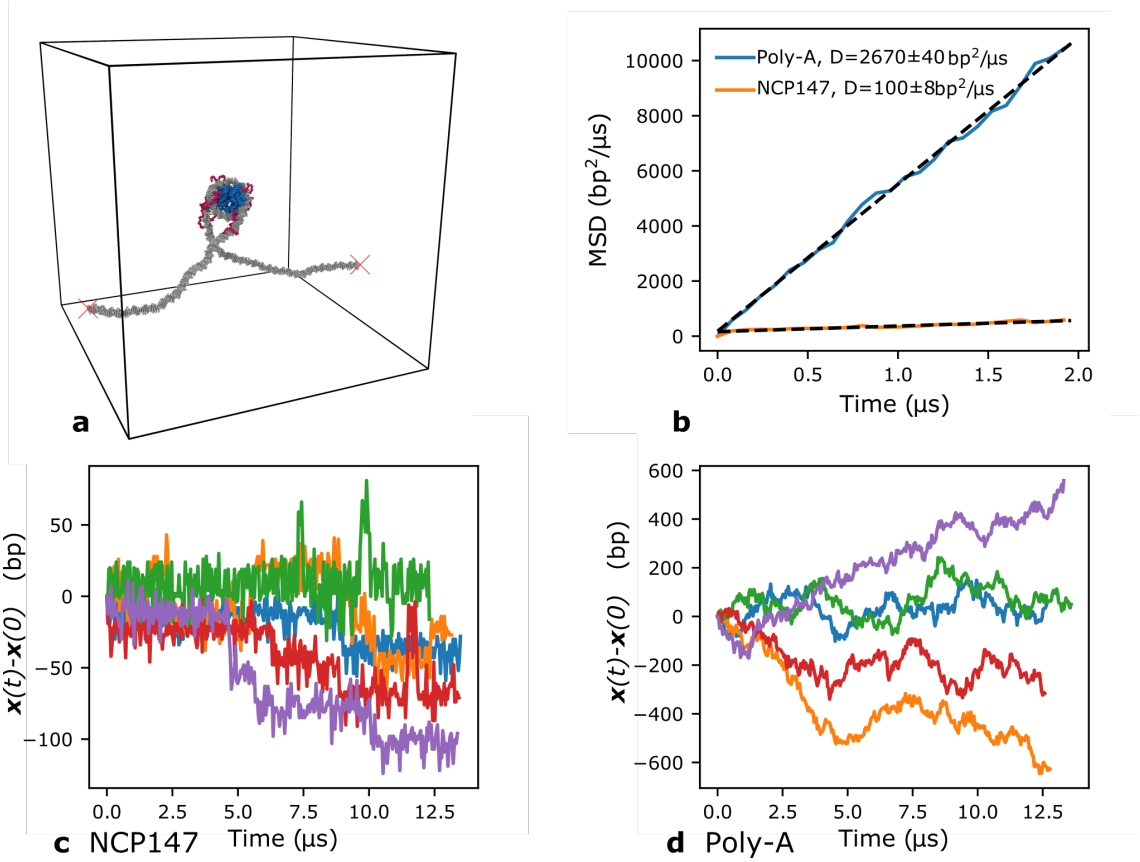

Figure 4.9: **DNA sequence dependent nucleosome sliding**. **(a)** Simulation setup, the box is periodic and, as indicated by the red crosses, the DNA strand is bonded to itself across the periodic boundary, effectively simulating an infinitely long strand of DNA. **(b)** MSD of a nucleosome on a poly-A sequence DNA strand (blue line) and a NCP147 sequence DNA strand (orange line). Both MSD lines have a straight line fit (black dashed lines) which have gradients equal to double the diffusion coefficient D which are listed in the legend. **(c-d)** Example time-series of the nucleosome position for 5 different trajectories for NCP147 and poly-A DNA sequences respectively, each different colored line represents a different simulation.

written as

$$\mathrm{MSD}(t) = 2Dt, \tag{4.5}$$

where D is the diffusion coefficient. Therefore D can be computed by fitting a straight line to a plot of MSD versus time. The computed MSDs from our simulation, corresponding linear fits, and values of D are plotted in figure 4.9a. Before plotting and fitting the MSD values were subsampled [178, 179], such that the time difference between data points was 40 ns (i.e. there are 50 data points for each MSD curve). The errors on the values of D are from the reported variance of the fitted parameters by the curve fitting function we used (scipy.optimize.curve_fit).

4.7.2 Results and discussion

Our simulations reveal that the diffusion coefficient for the poly-A sequence was approximately 26 times larger than for the NCP147 sequence (figure 4.9b). This shows that the nucleosomes are significantly more mobile on poly-A DNA, which

agrees with knowledge that poly-A DNA is typically nucleosome depleted [59] and with our results in section 4.5, where we found that the free energy for nucleosome unwrapping was lower for poly-A DNA compared to NCP147 DNA. We further agree with two other computational studies [175, 180], both using the coarse-grained model of the de Pablo group termed the 3PSN model; these studies find significant relationships between the DNA sequence and nucleosome sliding. Lequieu et al [175] found that the nucleosome sliding free energy landscape for the strong positioning '601' sequence exhibited much greater energy barriers than the free energy surface for the 'TTAGGG' repeat which positions nucleosomes poorly. The 601 and TTAGGG sequences are comparable to our NCP147 and poly-A sequences, respectively.

To probe the nucleosome sliding behavior in more detail, we examined the time-series of the nucleosome positions; examples of 5 randomly selected time-series are shown in figure 4.9c and d for NCP147 and poly-A respectively. We observed that the poly-A sequence exhibits frequent continuous movement consistent with that of a 1D random walk. This is in contrast to the NCP147 sequence, which exhibits rare, abrupt step like transitions. These observations are consistent with those of Niina et al [180], who observed the same trends for poly-CG and 601 sequences. These two different sliding modes, continuous sliding and step-like sliding, are similar to the proposed nucleosome sliding mechanisms of twist diffusion and loop propagation respectively. The noise in the time series, of approximately $\pm 10\,bp$, can be attributed to DNA breathing; this is, the unwrapping of the first 10 or so base-pair on each side of the nucleosome, leading to fluctuations in the value of the dyad location due to the method we use to estimate sliding explained earlier.

We note that the absolute value of the diffusion coefficients we report here are not that informative and should not be taken to represent the real timescales of in vivo nucleosome sliding. They are heavily dependent on our coarse-grained Langevin dynamics simulation method. However, the relative values between different DNA sequences are informative and allow us to conclude that the model accurately incorporates the sequence-dependent mechanical properties of DNA and predicts that these properties have important effects on the mobility of nucleosomes.

4.8 Orientation-dependent inter-nucleosome interactions

The nucleosome resembles a wedge-shape disk and has a contoured and irregular surface charge distribution. These features combined, imply that the strength of inter-nucleosome interactions have a dependence on the relative orientations of the nucleosomes. To quantify these differences, we computed the PMF of nucleosome pairs, at high (0.15M) and low (0.05M) monovalent salt, as a function of the inter-nucleosome center-to-center distance for three possible relative orientations: face-face, face-side, and side-side; these are illustrated in figure 4.10d.

4.8.1 Simulation methods

We computed a PMF for each of the three configurations using umbrella sampling simulations, as follows. An initial single nucleosome structure is equilibrated. Then, it is replicated and positioned such that the center-to-center vector between the pair

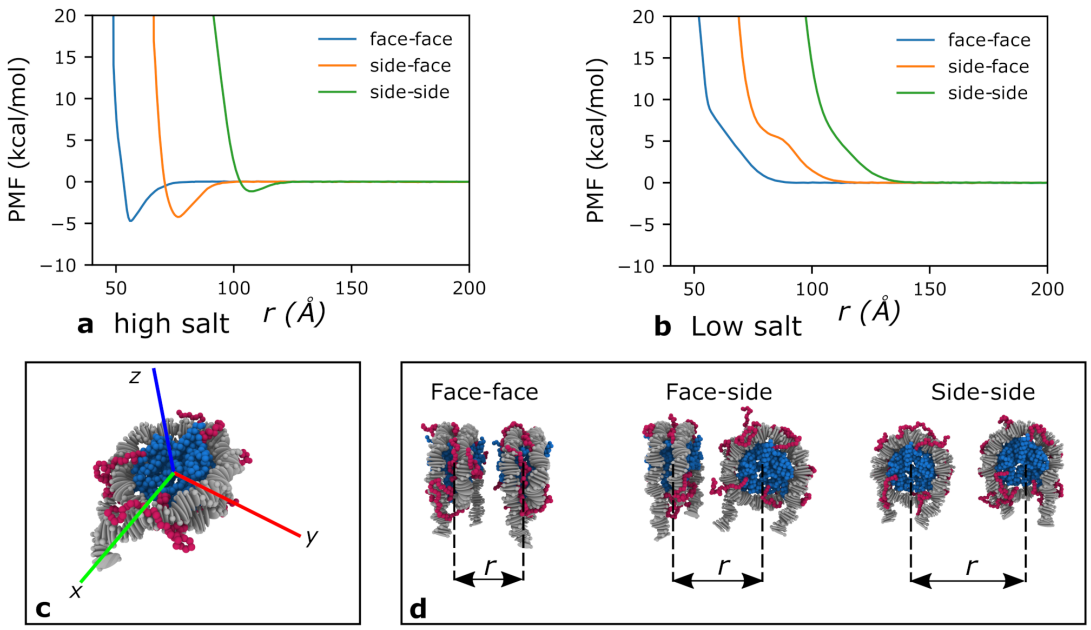

Figure 4.10: **Orientation dependent inter-nucleosome PMFs.** **(a)** PMF curves for the three inter-nucleosome orientation configurations at high salt (0.15 mol/L NaCl). **(b)** PMF curves for the three inter-nucleosome orientation configurations at low salt (0.05 mol/L NaCl). **(c)** Diagram showing the nucleosome axis. **(d)** Illustrations of the inter-nucleosome orientation configurations.

of nucleosomes is $\mathbf{r} = (0, 0, 200)$ Å. The nucleosomes are rotated into the desired orientations and held fixed using the COLVARS [162] `angleOrient` collective variable with a harmonic restraint of strength 1 kcal/mol/degrees^2 and a center of 0°. We use the LAMMPS `fix spring tether` command to restrain the nucleosomes to lie along the simulation box z-axis with harmonic restrains of strength 1 kcal/mol/Å^2 acting in the x and y directions only. Motion in the z direction is unaffected by this fix. The final collective variable is the distance between the nucleosome centers, r; this is the collective variable we use to perform umbrella sampling and compute the PMFs.

To prepare initial configurations for the umbrella sampling windows, r was varied using a SMD protocol from its initial value of 200 Å to its final value of 10 Å over 1 million timesteps with a force constant of 0.1 kcal/mol/Å^2. The range of r was split into 39 equally spaced windows, from 10 to 200. Each window was run for 10 million timesteps at the corresponding value of r with a force constant of 0.05 kcal/mol/Å^2. The orientational collective variables are kept with the same restraints as in the setup stage. Finally, the PMFs were computed from the window trajectories using WHAM [97].

4.8.2 Results and discussion

The computed PMFs are shown in figure 4.10a and b; these correspond to high and low monovalent salt conditions, respectively. We see that at high salt, the inter-nucleosome interactions are attractive, while at low salt they become purely repulsive. Although a nucleosome has a net negative charge (i.e. -137e), the screening effect of monovalent ions in solution at 0.15M is sufficiently large that the attractive

histone–DNA interactions overcome the DNA–DNA repulsion; we postulate that the histone tails are the main drivers of such effect, as they are highly positively charged and flexible, and their lengths (tens of amino-acids) are notably greater than the Debye screening length. As the concentration of counterions is decreased, the Debye-length increases, and as a consequence, the DNA-DNA repulsion begins to dominate, to the point where the histone tail-DNA attractive interactions get overtaken, and the overall nucleosome–nucleosome interaction becomes repulsive.

4.9 12-nucleosome chromatin structure: effects of nucleosome breathing

This section contains the main set of novel results on the structural behavior of chromatin that we obtained with our chemically-specific model. Importantly, these include the identification of the crucial, and previously unknown, impact of the spontaneous breathing motions of nucleosomes in enhancing the 'liquid-like' behavior of nucleosomes within compact chromatin.

Our work focuses on 12-nucleosome chromatin arrays with regular nucleosome repeat lengths (NRLs) of 165 bp and 195 bp because in vitro sedimentation coefficients are available for comparison. Compared to typical NRLs in nature, these represent short and long NRLs, respectively [181]. Our preliminary simulations carried out at 0.15M NaCl, showed that chromatin quickly condenses into configurations of high density that become kinetically trapped. This is due to the rugged energy landscape of the chemically-specific model: It has many degrees of freedom, many oppositely charged particles, and is a highly branched polymer. These tests revealed that an advanced sampling technique was necessary to sufficiently sample the phase space of chromatin at the relatively high resolution of our model. This observation is what led us to implement the Debye-length Hamiltonian replica exchange (HREMD) method that we described in section 3.8. When we initially considered using standard temperature replica exchange MD (TREMD), we found that because the number of replicas needed scales with the number of degrees of freedom in the system, for a 12N 165NRL system (i.e. $\sim 40,000$ particles) we would require ~ 80 replicas to span the temperature range 300-600 K. To overcome this challenge, we noted that the compaction of chromatin is modulated by the salt concentration, which for our model is completely incorporated in the value of λ_D in the electrostatic potential. Running preliminary simulations with $\lambda_D = 15.0$Å, we found that chromatin adopts an extended configuration, and that adequate sampling is readily achieved. Drawing a parallel with TREMD, using a larger λ_D value in our HREMD method has a similar effect to using a high temperature in TREMD, while the lowest value of $\lambda_D = 8$ Å (or a physiological salt concentration of 0.15 mol/L) corresponds to 300 K. We then tested exchanging Hamiltonian's between replicas and found that only 16 replicas were needed, with the values in table 4.4, to span the range $\lambda_D = 8$ to $\lambda_D = 15$ with exchange probabilities of approximately ~ 0.3. This is considerably less than the 80 replicas needed for TREMD. Furthermore, all the replicas are at physically meaningful salt concentrations, therefore while increasing sampling we are also investigating the salt-dependent behavior of the system.

4.9.1 Simulation methods

Using the HREMD method we ran simulations for four types of systems: chromatin with non-breathing nucleosomes and a NRL of either 165 bp or 195 bp; and chromatin with breathing nucleosomes and a NRL of either 165 bp or 195 bp. All of these simulations were run for more than 100 million timesteps with an HREMD exchange frequency of 10,000 timesteps. Each set of exchanges either attempts to exchange replicas {1-2,3-4...15-16} or {2-3,4-5...14-15}, with each set picked with a 50% probability. Coordinate snapshots were recorded to the trajectories every 100,000 timesteps. The last 50 million timesteps were used for analysis.

For qualitative assessment of the chromatin dynamics, we took equilibrated structures for both breathing and non-breathing chromatin at 0.1 mol/L and ran them using standard molecular dynamics for ∼2 µs.

Replica	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
λ_D (Å)	8.00	8.34	8.70	9.07	9.46	9.86	10.29	10.73	11.19	11.67	12.16	12.69	13.23	13.79	14.38	15.00

Table 4.4: HREMD Debye-length (λ_D) values.

4.9.2 Analysis methods

4.9.2.1 Sedimentation coefficients

The sedimentation coefficient of a molecule is a measure of its compactness. It behaves similarly to the inverse of the radius of gyration, and can be easily measured experimentally (for more details on sedimentation coefficients see section A.3.1). We computed the sedimentation coefficients of chromatin using the HullRad method [182], which we describe in section A.3.

4.9.2.2 Nucleosome valency

Because the valency of a biomolecule is intimately linked to its ability to undergo liquid–liquid phase separation (which we explore later in this thesis), we quantify the nucleosome valency in our simulations. We define the valency of a nucleosome as the average number of other nucleosomes that each nucleosome is in contact with via the following relation:

$$\text{Valency} = \frac{1}{N_t N_n} \sum_{t}^{N_t} \sum_{i,j \neq i}^{N_n} C_{i,j}(t), \tag{4.6}$$

$$C_{ij}(t) = \begin{cases} 1, & \text{if nucleosomes i and j are in contact,} \\ 0, & \text{otherwise,} \end{cases} \tag{4.7}$$

where i and j sum over the $N_n = 12$ nucleosomes in each frame, t sums over all N_t frames in the trajectory, and two nucleosomes are classed as being "in contact" if the distance between their centres of mass is less than 110 Å.

4.9.2.3 Amount of unwrapped DNA

The amount of unwrapped DNA is defined as the average number of DNA base-pairs that unwrap from the nucleosomes per nucloeosome (relative to a nucleosome having 147 base-pairs). Full details of how we define nucleosomal and linker DNA in the breathing chromatin simulations is given in section A.4.

4.9.2.4 Inter-nucleosome interactions

The inter-nucleosome interactions are a measure of the contact frequency between nucleosomes that are k^{th} neighbours of each other. The interactions are then categorised into either face–face, face–side, or side–side, which are the same categories as in section 4.8. Full details of the calculation are given in section A.5.

4.9.2.5 Molecular-level inter-nucleosome contacts

The molecular-level inter-nucleosome contacts are defined as the inter-nucleosome contacts mediated specifically by each of the different amino acid beads and DNA beads (e.g., it can reveal which specific protein or DNA regions sustain chromatin compaction). The full details of the calculation are given in section A.6.

4.9.3 Results

Our simulations demonstrate that nucleosome breathing has a significant effect on the structure of chromatin with short NRLs (i.e. 165 bp). For the majority of the salt range studied, the values of the sedimentation coefficients in figure 4.11 differ for 165 NRL chromatin with non-breathing versus breathing nucleosomes. With respect to their non-breathing counterparts, breathing nucleosomes enhance the compaction of 165 bp chromatin at higher salt, but favor more extended configurations at low salt. For chromatin with an NRL of 195 bp (figure 4.11b), the difference in the observed sedimentation coefficients is minimal; this is because the relatively long linkers of the 195 bp systems intrinsically introduce structural heterogeneity. The histograms in 4.11 show that for 0.15 mol/L NaCl, our measured sedimentation coefficients are in quantitative agreement with values from experiments [183]. Moreover, the progressive de-compaction of chromatin with decreasing salt is in qualitative agreement with the experimental trend [183].

Figure 4.12a and c show the nucleosome valency for 165 NRL and 195 NRL, respectively. The marked difference between chromatin with non-breathing and breathing nucleosomes for the shorter NRL is notable here. At higher salt, the breathing nucleosomes exhibit an increased valency; that is, breathing nucleosomes arrange irregularly within chromatin, and lead to denser structures by forming more connections with their neighbours (illustrated in the valency=5 cartoon). In contrast, non-breathing nucleosomes arrange into an ordered and regular 'zig-zag' fiber, where they typically only come into contact with the nucleosomes positioned $k = 2$ away from them (illustrated in the valency = 2 cartoon and figure 4.13).

To quantify the extent of nucleosome breathing, we computed the number of DNA base-pairs that unwrap per nucleosome across the simulation (plotted in figure 4.12b and c). We observed that as the salt concentration increases, the the

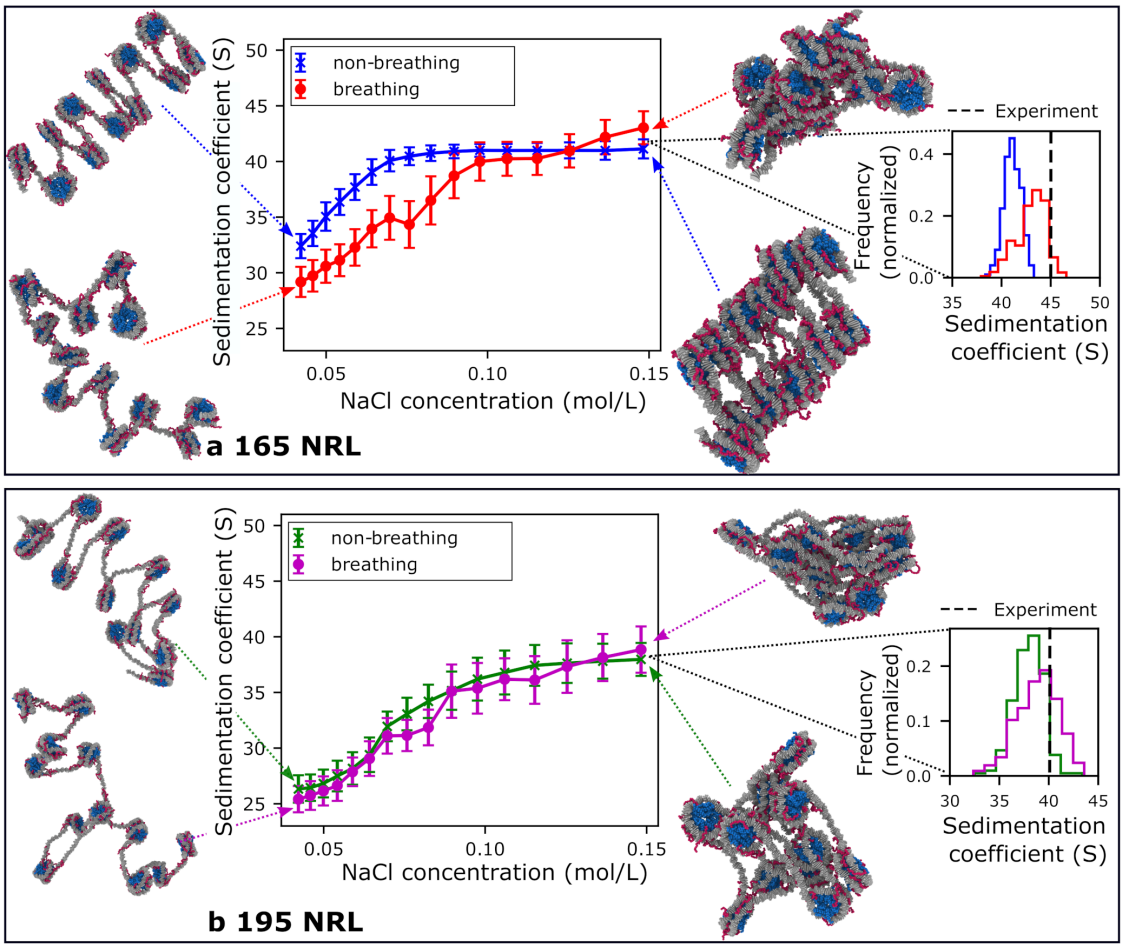

Figure 4.11: **Sedimentation coefficients of 12-nucleosome chromatin. (a)** Sedimentation coefficients for 165 NRL chromatin for breathing (red points) and non-breathing (blue points) nucleosomes. **(b)** Sedimentation coefficients for 195 NRL chromatin for breathing (magenta points) and non-breathing (green points) nucleosomes. The errorbars on the both plots are standard-deviations. The histograms attached to each plot show the full distributions for the 0.15 mol/L data points. The vertical dashed lines are the experimental values for 12-nucleosome chromatin at similar NRLs from [183]. The simulation snapshots illustrate typical chromatin conformations at the corresponding labelled parts of the graphs.

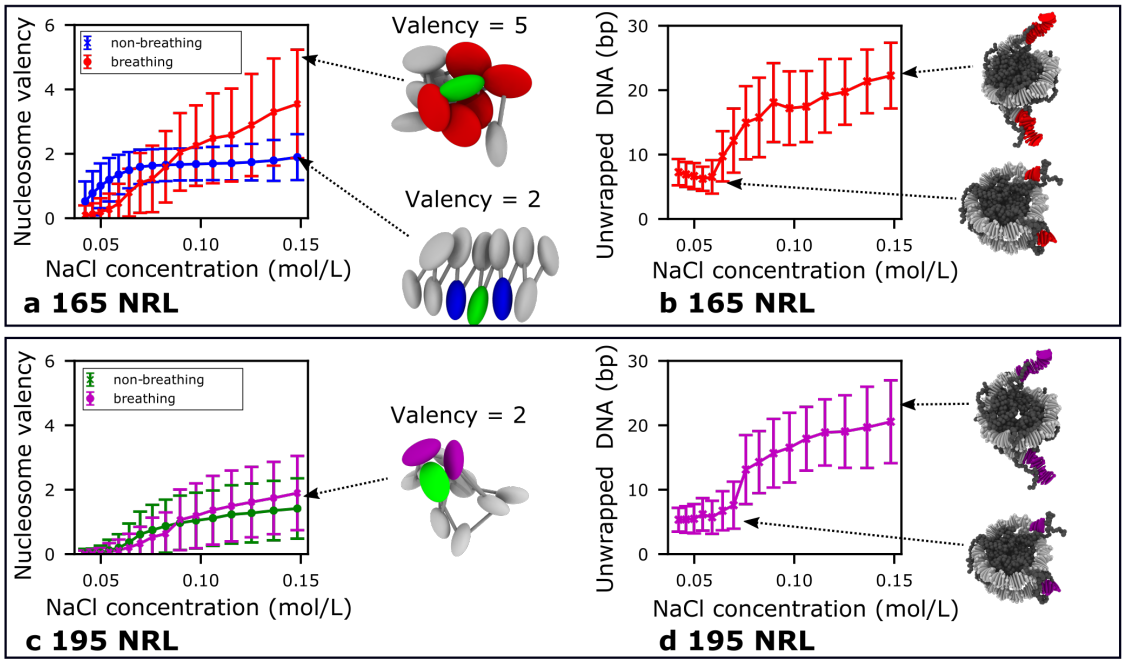

Figure 4.12: **Nucleosome valency and amount of unwrapped DNA. (a)** Nucleosome valency for 165 NRL. **(b)** Amount for unwrapped DNA per nucleosome for 165 NRL. **(c)** Nucleosome valency for 195 NRL. **(d)** Amount of unwrapped DNA per nucleosome for 195 NRL. The error-bars on all plots are standard deviations. The cartoon images in **a** and **c** illustrate example configurations with the indicated valency. The nucleosome images in **b** and **d** indicate the amount of DNA that unwraps from the nucleosome at the corresponding parts of the graphs.

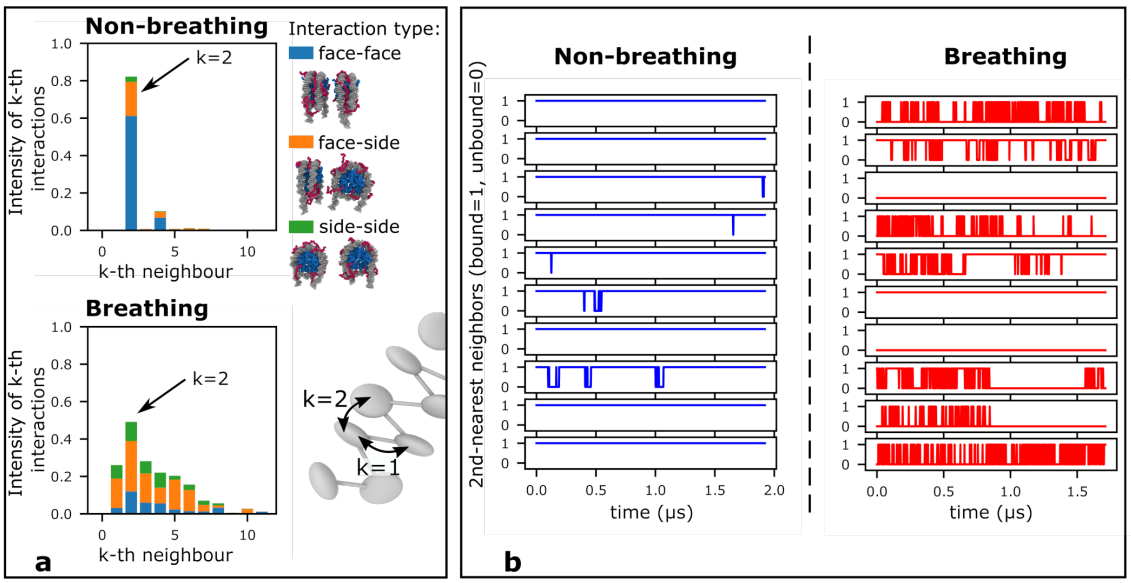

Figure 4.13: **Inter-nucleosome interactions. (a)** Intensity of k^{th} neighbour interactions categorised by relative orientation for non-breathing (top panel) and breathing (lower panel) chromatin at 0.15 mol/L. **(b)** Time series of $k = 2$ interactions. The 10 plots correspond to the 10 possible $k = 2$ interactions in 12-nucleosome chromatin.

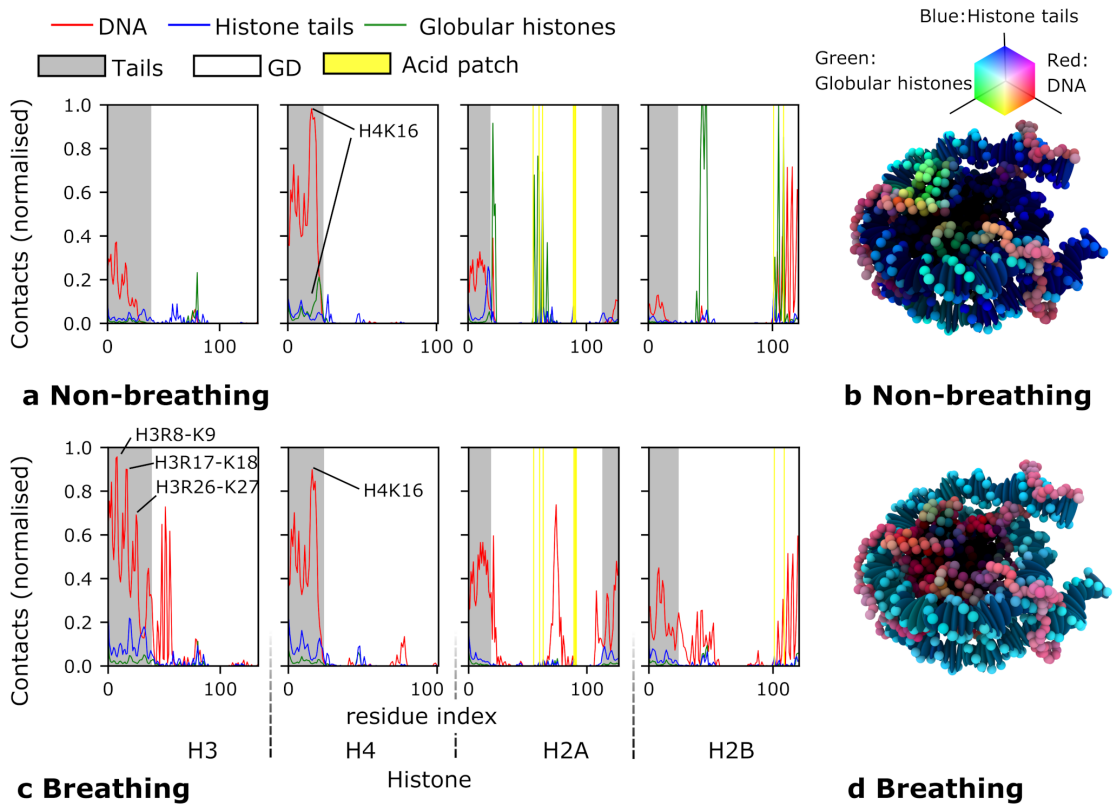

Figure 4.14: **Molecular-level inter-nucleosome contacts. (a, c)** normalized fraction of nucleosome-nucleosome molecular-level contacts within compact chromatin for non-breathing and breathing nucleosomes respectively. The horizontal axis runs across all histone protein residues within the nucleosome (H3, H4, H2A, and H2B). The gray shaded areas are the histone tail regions, the white areas the globular domains, and the vertical yellow lines indicate the acidic patch residues. The contacts are broken down by type protein-DNA (red), protein-histone tail (blue), and protein-globular domain (green). **(b, d)** Visualisation of the contacts. Each residue and DNA base-pair are colored according to a RGB value that is obtained by combining the values of the red, green, and blue lines in **a** and **c**, with a logarithmic scaling.

number of unwrapped bases-pairs increases from 7±2bp to 22±5bp; this occurs because the DNA-histone electrostatic attraction weakens under conditions of higher monovalent salt screening. The enhancement of nucleosome breathing at increasing salt is consistent with in vitro single-molecule experiments where nucleosomes are considerably unwrapped (∼30%) at 0.1 mol/L NaCl but minimally unwrapped (∼10%) at 0.02 mol/L NaCl [184].

Focusing further on the 165 NRL simulations, we computed the intensity of the orientation-dependent interactions between nucleosomes that are $k-th$ neighbours of each other. The top panel of figure 4.13 shows that, for chromatin with non-breathing nucleosomes, the $k = 2$ interactions are dominant, predominately in the face–face orientation; this is the hallmark of the regularly stacked 'zig-zag' structure. In contrast, the lower panel of figure 4.13 shows that for chromatin with breathing nucleosomes, inter-nucleosome interactions are highly diverse: i.e., they occur between a wide-range of nucleosome pairs and in many different inter-nucleosome orientations.

To qualitatively probe the dynamic nature of the inter-nucleosome interactions, we next ran standard MD simulations (i.e. without enhanced sampling) for our 165 NRL breathing system and a 165 NRL non-breathing system at 0.1 mol/L NaCl, and measured the time series of the $k = 2$ interactions. These are plotted in figure 4.13b, where the ten plots for each system correspond to the 10 possible $k = 2$ nucleosome pairs. The interactions in the non-breathing systems are longer lived and mostly constant for the duration of the simulation (i.e., once nucleosomes stack in a zigzag fiber, they rarely unstack). In contrast, the interactions among breathing nucleosomes exhibit frequent fluctuations.

The ability of our model to resolve the motions of individual amino acids and DNA base-pairs within compact chromatin, allows us to examine the precise contributions of each of these species in directing chromatin organisation. Specifically we computed the fraction of time each amino acid and DNA base-pair in a given nucleosome mediate inter-nucleosome interactions. The interactions were categorised into three main groups: DNA, globular regions, and histone tails. This analysis, plotted in figure 4.14, revealed strikingly different interaction patterns between the non-breathing and breathing chromatin.

The face-face stacking of non-breathing chromatin is shown by the green line peaks in panel H2A and H2B. These show interactions where the amino acids within globular domains are in frequent contact with other amino acids also belonging to globular domains. The important acid patch region — a cluster of negatively charged amino acids at the centre of the nucleosome face, known to mediate many nucleosome-protein interactions and nucleosome-nucleosome interactions within zigzag fibers — are included in these regions. For breathing chromatin, the interactions with globular histones are reduced, and instead the interactions with DNA are favored (e.g. the red line peak present in figure 4.14 at position 80 in H2A is not present in figure 4.14). The H4 tail mediated interactions are prominent in both chromatin with breathing and non-breathing nucleosomes. Among the H4 residues, H4K16 forms the most frequent contacts with both the DNA and the acid patch; this is due to its positive charge (K = Lysine which has +1 charge) and its electrostatic attraction with the negatively charged DNA, and more moderately with the acid patch region in the case of the non-breathing chromatin.Strikingly, this dominance of the H4K16 residue is in agreement with the well-known de-compaction

triggered by H4K16 acetylation [80, 185], and the observation that reversible acetylation of H4K16—one of the most frequent post-translational modifications across organisms—has diverse functional implications [186]. The H3 tail interactions with DNA are also significant, this is due to the abundance of the positively charged arginine and lysine residues (labelled on figure 4.13, R=arginine, K=lysine). The importance of histone tail–DNA electrostatic interactions is supported by experiments demonstrating that chromatin with tail-less nucleosomes fails to condense [185].

4.9.4 Discussion

Our simulations predict that nucleosome breathing sensitively affects the structural behavior of chromatin with short linker DNAs. In particular, 165 NRL chromatin with non-breathing nucleosomes adopts a regular 30nm 'zig-zag' like conformations, in agreement with short NRL 12-nucleosome chromatin structures observed in cryo-EM experiments [15], and the 167 NRL tetranucleosome crystallographic structure [14]. This is in striking contrast to the 165 NRL chromatin with breathing nucleosomes which exhibits an irregular, fluctuating, 'liquid-like' structure, as has been postulated by Maeshima and collaborators [19]. Importantly, this irregular behavior of chromatin is in qualitative agreement with the disordered organisation of nucleosomes inside cells observed with super-resolution microscopy[23], and chromEMT experiments [50]. The liquid-like behavior of nucleosomes within chromatin emerges as a consequence of the DNA being able to unwrap from the nucleosome core; this in turn weakens the strong torsional restraints imposed by the linker DNA, and allows nucleosome–nucleosome interactions to occur with more diverse orientations, furthermore, the system has increased entropy. For the 195 NRL simulations we found that the difference between breathing and non-breathing chromatin was minimal, this is because the longer linker DNA can intrinsically create the nucleosome-nucleosome oritentational heterogeneity that sustains chromatin's liquid-like behavior.

The modulation of nucleosome breathing with salt and its impact on chromatin structure may help explain why ordered and disordered structural chromatin models have been derived from in vitro and in vivo data, respectively. In vitro experiments typically use low salt conditions and highly regular reconstituted chromatin arrays i.e. with strong nucleosome positioning sequences and uniform NRLs [11, 18]. Experiments and our simulations demonstrate that low salt conditions hinder nucleosome breathing while higher physiological salt conditions (above ∼0.1 mol/L of NaCl) promote it [184]. Our work further reveals that significant nucleosome breathing at physiological salt favors the liquid-like behavior of chromatin even in artificially homogeneous chromatin (i.e., with the uniform DNA linker lengths and DNA sequences we have used).

4.9.4.1 Quantitative vs qualitative salt dependent behavior

The compaction of chromatin predicted by our model is in quantitative agreement with sedimentation coefficient experiments at 0.15mol/L monovalent salt concentration, this is shown in the sedimentation coefficient histograms in figure 4.11. In addition, the force-induced unwrapping behavior of single-nucleosomes stemming from our model is in quantititave agreement with single-molecule force-spectroscopy experiments at 0.15 mol/L NaCl, as described in section 4.5. Our agreement with

experiments precisely at a value of 0.15 mol/L NaCl is significant; this is the salt concentration that we are most interested in, because it approximates well the solvent conditions in vivo (referred to as physiological salt concentration) and is used widely in vitro experiments. In our model this salt condition corresponds to using a Debye-length of 8Å in the screened Coulomb interaction (via equation 3.12). When we increase the Debye-length in our model to 15Å, we observe that chromatin decondenses; this enables us to use the Debye-length Hamiltonian replica exchange method to sample the 0.15 mol/L chromatin configurations.

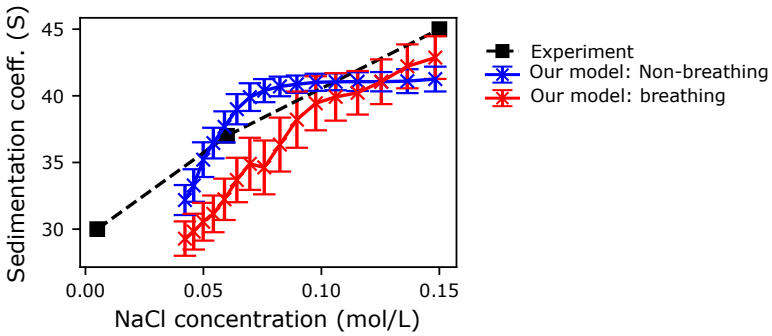

Figure 4.15: **Comparison of chromatin salt dependent compaction for 12-nucleosome 165 NRL between our model and the experimental values for in vitro chromatin from Correll et al [183]** . The blue and red data points are the same as those in figure 4.11a. The black points (labelled experiment in the legend) are from Ref. [183], The lowest salt concentration data point is at 0.005 mol/L.

A Debye-length of 15Å corresponds to a NaCl concentration of 0.042 mol/L. The sedimentation coefficient we observe of ∼30 S at this salt concentration is most similar to the sedimentation coefficient Correll et al [183] observe for 0.005 mol.L, this is shown in figure 4.15. This highlights that for lower salt concentrations our model captures the salt dependent behavior qualitatively — i.e. as the salt concentration is decreased, chromatin becomes less compact. However, our model predictions of chromatin compaction are not in quantitative agreement over the whole salt range, only for the larger values of approximately greater than 0.1 mol/L. At this stage we are not too concerned with this because we primarily use the HREMD method to sample the 0.15 mol/L chromatin configurations and we are mostly interested in the relative differences between chromatin behavior for different conditions within our model framework, the absolute values of the salt concentration are less important. In future work, not in this thesis, we plan to investigate this property of the model more and seek to achieve quantitative salt-dependent behavior.

4.10 12-nucleosome chromatin: effects of H1 linker histone

Histone H1, also called linker histone (LH), is a histone protein present in eukaryotic cells. It binds to nucleosomes, typically in or around the dyad position [64] — as is pictured in figure 3.6. The combined unit of nucleosome plus bound linker histone is called a chromatosome [187, 188]. Linker histone proteins have a tripar-

tite structure: a short N-terminal domain (20 amino-acids), a structured globular domain (70 amino-acids), and a long positively charged C-terminal domain (100 amino-acids) [64]. The role of linker histone on the structure of chromatin is still not fully understood, previous studies have found that binding of linker histone to chromatin stabilizes regular folding [15, 16, 189], while others have found that linker histone containing chromatin exhibits irregular folding [23, 50, 64]. We use our model to investigate the effects of linker histone on short 165 NRL 12-nucleosome chromatin, for which we saw the most significant difference between breathing and non-breathing chromatin.

4.10.1 Simulation methods

Our preliminary simulations indicated that 165 NRL chromatin with LH remains compact even with a Debye-length of 15.0Å. This means the HREMD method used in the previous section will not be appropriate without extra tuning. Taking a different approach we first narrow the phase-space by considering non-breathing chromatin only, for LH simulations this means we include the LH globular domain, the nucleosomal DNA, and the histone core globular domain as one unit connected by a GNM. We then further simplify the model by replacing the histone core GNM and the linker histone GNM with composite rigid bodies using the LAMMPS command `fix rigid`. This reduces the degrees of freedom from 54000 to 25000 which using TREMD requires 64 replicas to span 300K-600K with an acceptance probability of approximately 0.4. We simulated 12-nucleosome 165 NRL non-breathing chromatin with LH at two salt concentrations, 0.15 M and 0.05 M using TREMD with rigid-body nucleosome cores and LH globular domains. The simulations were run for approximately 100 million timesteps, TREMD exchanges were attempted every 1000 timesteps, and simulation snapshots were recorded every 100,000 timesteps.

4.10.2 Analysis methods

We computed the sedimentation coefficients and the inter-nucleosome interactions using the same methods as in section 4.9.

4.10.3 Results and Discussion

The sedimentation coefficients and inter-nucleosome interactions are plotted in figure 4.16a and b respectively. We observed that with LH the chromatin compaction increased (larger sedimentation coefficients). Furthermore, the zig-zag ladder structure we observe for 12-N 165 NRL non-breathing is destabilized, this is evident from the inter-nucleosome interactions which do not exhibit the dominant k=2 face-face interactions typical of the zig-zag ladder. This effect of LH, in destabilizing the zig-zag ladder structure of short NRL chromatin, is in agreement with the EM experiments of Routh et al [190]. It is also consistent with our simulations showing that the C-terminal tail of H1 remains disordered upon nucleosome binding, and that such disorder enhances the linker DNA fluctuations and destabilizes 30-nm fiber folding [64]. Finally, it also agrees with experiments showing that the C-terminal domain of H1 remains disordered when bound to DNA [191] or to a long negatively charged disordered protein [192], and the hypothesis derived from there that LH behaves as a liquid-like glue for chromatin [193].

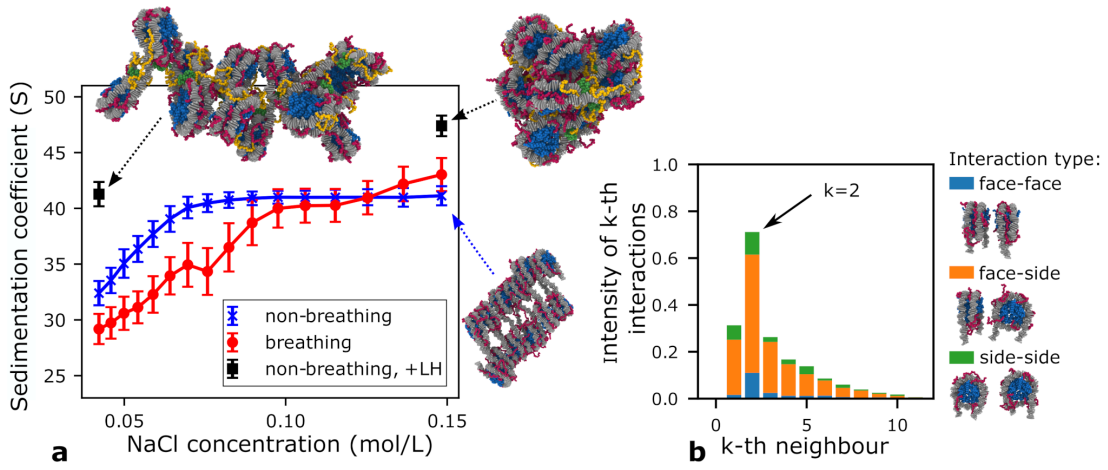

Figure 4.16: **Effects of H1 linker histone (LH) on the structure of 12-nucleosome 165 NRL chromatin (non-breathing).** **(a)** Sedimentation coefficents of 12-nucleosome 165NRL chromatin as a function of the salt concentration. The blue and red points are the same as figure 4.11a, the black points are for chromatin with linker-histone. **(b)** The k^{th} neighbour interactions for chromatin with linker histone at 0.15 mol/L. The interactions are categorised into either: face-face, face-side, side-side.

Chapter 5

Minimal model

In this chapter we explain the development of the minimal model (level 3 in the multiscale hierarchy). The development of this model is part of the publication [121].

Contents

5.1 Minimal representation

A key challenge in the field of chromatin organization is to describe nucleosome behavior within domains that are larger than the sizes of genes (tens to hundreds of kilobases), and thus, have functional relevance. Thus, we further coarse-grained our model with the aim of being able to simulate systems with hundreds of nucleosomes. The essential information about chromatin we wish to keep are the details of the orientation dependent nucleosome-nucleosome interactions, the ability to resolve the effects of DNA breathing, and correct inclusion of the semi-flexible/twistable nature of the linker DNA. These requirements led us to propose a model which represents the nucleosome histone core with a single ellipsoidal bead, and the DNA with finite-sized spheres that each represent 5 base-pairs. The reason for choosing 5 bp, rather than a larger number, is to do with the excluded volume size of the DNA. The diameter of the DNA double helix is ≈ 20Å, thus to prevent the possibility of unphysical chain crossing with a bead-spring style polymer, the bond length between adjacent beads must be smaller than the excluded volume size (which has a minimum possible value of 20Å). This is satisfied by 5 base-pairs per bead because 5 times the rise of DNA (3.3Å) gives a bond length of 16.5Å.

Using our knowledge of nucleosome geometry, we set the shape of the core ellipsoids to (28Å $\times$ 28Å $\times$ 20Å) and the DNA ellipsoids to (12Å $\times$ 12Å $\times$ 12Å). The DNA beads are spherical, which for implementation reasons are ellipsoids with all radii equal. The DNA beads are categorized into linker DNA or nucleosomal DNA,

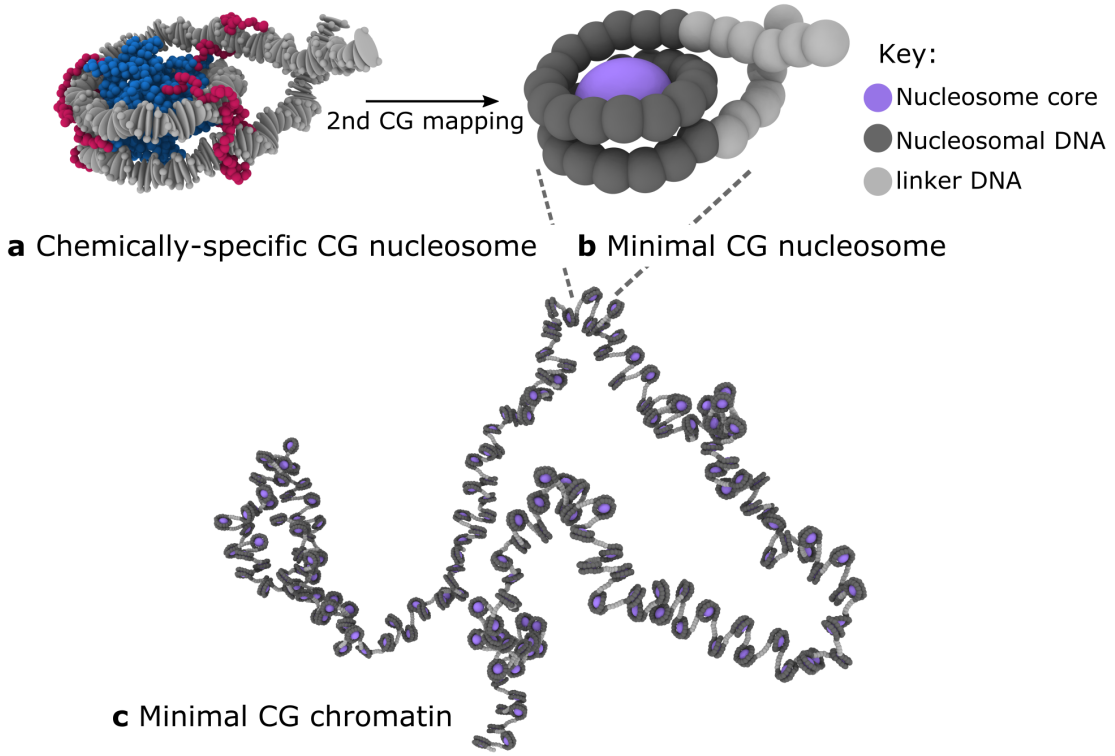

Figure 5.1: **Minimal model (level 3 in the multiscale hierarchy).** **(a)** Chemically specific model nucleosome ∼1000 particles. **(b)** Minimal model nucleosome ∼40 particles. **(c)** Minimal model 200-nucleosome chromatin.

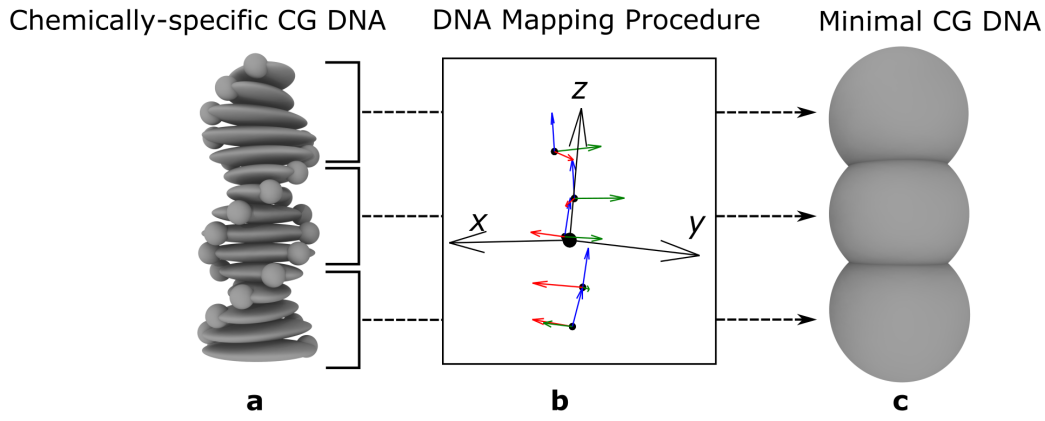

Figure 5.2: **Mapping from chemically-specific model DNA to minimal model DNA.** **(a)** 15 base-pairs of chemically-specific model DNA. **(b)** Mapping procedure, the positions and orientations of the 5 DNA base-pairs are averaged to give the position and orientation of the minimal model DNA bead. **(c)** 3 minimal model DNA beads which represent 15 base-pairs.

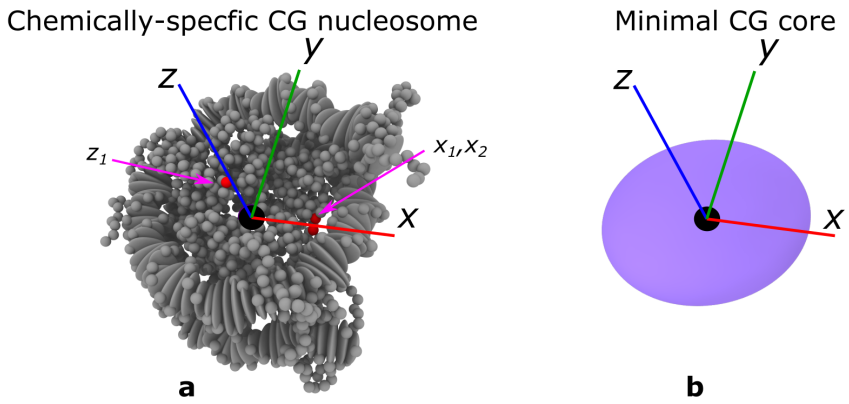

Figure 5.3: **Mapping from chemically-specific model histone core to minimal model core bead.** **(a)** Chemically-specific nucleosome. As described in the text, the labelled protein beads x_1, x_2, z_1, and z_2 are used to construct the nucleosome axis. **(b)** Minimal model core bead, represents the core histone proteins.

where the difference is that nucleosomal DNA is bound to its respective core bead. To be consistent with the chemically-specific model, we set the masses of the beads to the approximate values they represent, as listed in table 5.1.

Particle type	Atom ID	Shape (Å)	Mass (g/mol)
Core	1	$(28 \times 28 \times 20)$	100000
Linker DNA	2	$(12 \times 12 \times 12)$	3250
Nucleosomal DNA	3	$(12 \times 12 \times 12)$	3250

Table 5.1: Minimal model particle properties.

5.2 Mapping procedures

To create the initial structures we map directly from the chemically-specific model structures using the DNA and core mapping procedures explained below.

5.2.1 DNA mapping

For the DNA we consecutively group the chemically-specific DNA beads into sets of 5, each set has the position vectors $\mathbf{r}_i$ and quaternions $\underline{q}_i$ where $i = 1, 2, 3, 4, 5$. The position of the minimal DNA bead $\mathbf{r}_M$ is average position:

$$\mathbf{r}_M = \frac{1}{5}\sum_{i=1}^{5}\mathbf{r}_i, \tag{5.1}$$

and the orientation is the normalized average quaternion

$$\underline{q}'_M = \frac{1}{5}\sum_{i=1}^{5}\underline{q}_i, \tag{5.2}$$

$$\underline{q}_M = \underline{q}'_M/|\underline{q}'_M|. \tag{5.3}$$

We note than in general taking the mean of a quaternion in this way may not always be accurate [194], but in our use case it is sufficient because the quaternions we average over represent similar orientations. Additionally, we account for quaternion double cover (as $\underline{q}$ represents the same rotation as -$\underline{q}$), by setting $\underline{q}_i$ to $-\underline{q}_i$ if $\underline{q}_1 \cdot \underline{q}_i < 0$, before taking the mean.

5.2.2 Histone core mapping

To fit the minimal model core bead, we set its position as the center of mass of the histone core globular domain and we set the orientation by constructing the nucleosome's orientation axis unit vectors in a consistent manner using specific amino acid beads. We compute the center of mass $\mathbf{r}_{COM}$ and record the positions of the four specific amino acid beads $\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}_1$, and $\mathbf{z}_2$ (labeled in 5.3) which have particle indexes of 113, 600, 466, 953 respectively. The x-axis is given by the vector

$$\mathbf{x} = (\mathbf{x}_1 + \mathbf{x}_2)/2 - \mathbf{r}_{COM}, \tag{5.4}$$

the approximate z-axis is given by

$$\mathbf{z}' = \mathbf{z}_1 - \mathbf{z}_2, \tag{5.5}$$

the y-axis is given by

$$\mathbf{y} = -\mathbf{x} \times \mathbf{z}', \tag{5.6}$$

now we compute the z-axis that is orthogonal x and y

$$\mathbf{z} = \mathbf{x} \times \mathbf{y}. \tag{5.7}$$

We normalize these vectors and can convert them back into the quaternion representation which is the orientation of the minimal core bead.

5.3 Breathing and non-breathing nucleosomes

The difference between non-breathing and breathing nucleosomes in the minimal model is only the initial structures that are used to build them; that is, both models contain DNA that is classified as either nucleosomal DNA, which is rigidly fixed to the nucleosome core, or linker DNA, which is unconstrained. For the breathing structures the nucleosomes are in configurations where some of the DNA is unwrapped (i.e. no longer part of the nucleosome rigid body). These configurations are taken from the thermodynamic probability distribution of nucleosome breathing states at the simulated salt concentration, i.e. from our chemically-specific model simulations in section 4.9. The reason we do this is to significantly reduce the degrees of freedom in the system. The nucleosomes are still able to dynamically unwrap to the levels expected, i.e figure 4.12b and d, but the nucleosomes are unable to slide. This is not too important as we found that the rate of nucleosome sliding, for a strong positioning sequence, is slow relative to the timescales we simulate (section 4.7).

5.4 Generating initial structures

We have two methods for generating initial structures for our simulations. The first is to directly map equilibrium structures from the 12-nucleosome chemically-specific model simulations to the minimal model representation. This is shown in figure 5.4, where the differences between minimal model breathing and non-breathing nucleosome becomes clear.

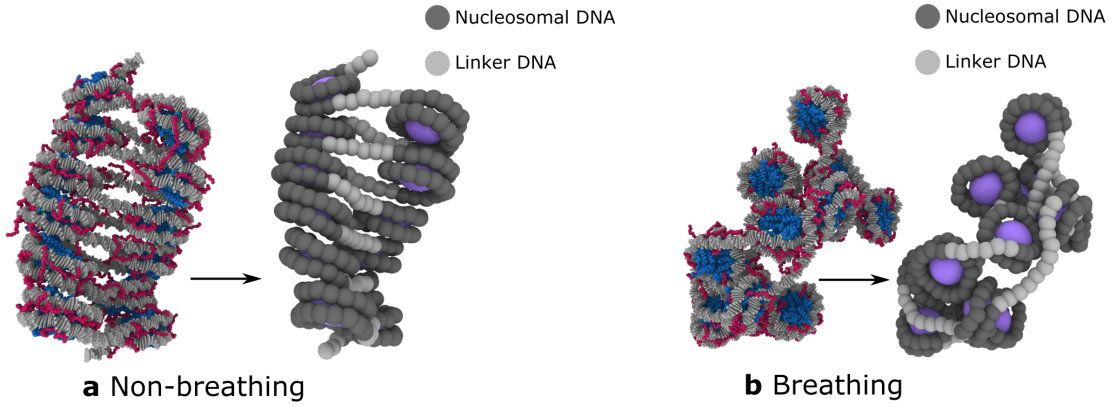

Figure 5.4: **Generating minimal model initial structures version 1**. 12-nucleosome structures are directly mapped from equilibrium chemically-specific model structures. The important difference between the non-breathing (**a**) and the breathing (**b**) chromatin is the amount of linker DNA, at higher salt the non-breathing chromatin has significantly more.

The second method, figure 5.5, is very similar to how we generate the initial structures for the chemically-specific model. We begin with the full all-atom reference nucleosome structure, then we perform the chemically-specific model CG mapping, directly followed by the minimal model CG mapping. We now have a minimal model nucleosome. To create chromatin of length N nucleosomes we simply connect N nucleosomes together. Then, for the breathing chromatin we redefine the nucleosomal DNA classification by unwrapping X amount of DNA from the ends of the nucleosomes where X is a random number, different for each nucleosome, sampled from a normal distribution with mean and standard deviation taken from figure 4.12b.

5.5 Potential energy function

Following our methodology of "bottom-up" coarse-graining, we use potentials that can be fitted from our chemically-specific model simulations. The potential energy function that we use is

$$E = E_{\text{Minimal-LJ}} + E_{\text{Minimal-RBP}} + E_{\text{Anisotropic}}. \tag{5.8}$$

E_{LJ} is a short range pairwise interaction that occurs between all beads, $E_{\text{Minimal-RBP}}$ is a bonding potential between DNA, and $E_{\text{ansiotropic}}$ is a short range orientation dependent attractive potential that approximates DNA binding to the nucleosome core.

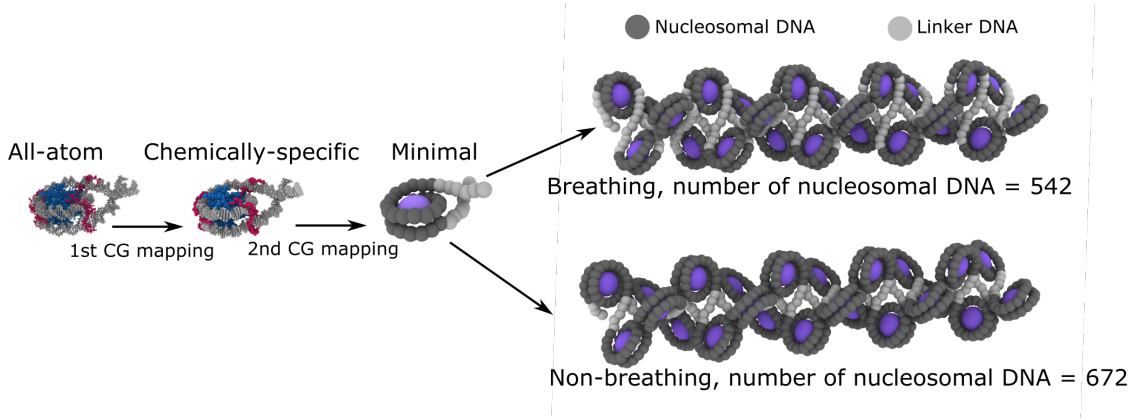

Figure 5.5: **Generating minimal model structures version 2**. Minimal model nucleosomes are directly created from the all-atom nucleosome reference structure. We connect nucleosomes together in the same manner as section 3.6. For breathing chromatin some nucleosomal DNA is unwrapped (recategorized as linker DNA), following the salt dependent DNA unwrapping behavior we observe in the chemically specific model (figure: 4.12b).

5.5.1 Bonded terms

For the bonded term we use a rigid base-pair like potential which, instead of describing the interactions between each base-pair, describes interactions between beads which each represent 5 base-pairs. The functional form is the same as the standard RBP potential

$$E_{\text{Minimal-RBP}} = \frac{1}{2}\Delta\boldsymbol{\phi}\mathbf{K}\Delta\boldsymbol{\phi}, \tag{5.9}$$

$$\Delta\boldsymbol{\phi} = (\boldsymbol{\phi} - \boldsymbol{\phi}_0), \tag{5.10}$$

where ϕ are the helical parameters describing the relative positions and orientations of the two minimal model DNA beads in the bond. At this level of resolution the effects of the DNA sequence start to average out therefore we have the same parameter set ($\mathbf{K}$, $\boldsymbol{\phi}_0$) for all bonds.

5.5.1.1 Fitting parameters

We fit the parameters for the minimal RBP potential by following the same procedure that can be used to fit standard RBP potential parameters from atomistic DNA simulations. That is, if we can generate a set of values of ϕ_i (where $i = \{$shift, slide, rise, tilt, roll, twist$\}$) sampled from the canonical distribution then ϕ_{0i} are found as the mean

$$\phi_{0i} = \langle\phi_i\rangle, \tag{5.11}$$

and $\mathbf{K}$ is found by inverting the covariance matrix

$$\mathbf{K} = k_B T \mathbf{C}^{-1}, \tag{5.12}$$

$$C_{ij} = \langle(\phi_i - \langle\phi_i\rangle)(\phi_j - \langle\phi_j\rangle)\rangle. \tag{5.13}$$

To generate the set of ϕ we ran simulations of 200 bp strands of DNA using the chemically-specific model. We did 10 repeats with a different random DNA sequence for 100 million timesteps, outputting every 10,000 timesteps. Each frame from the trajectory was converted into the minimal representation and the helical parameters

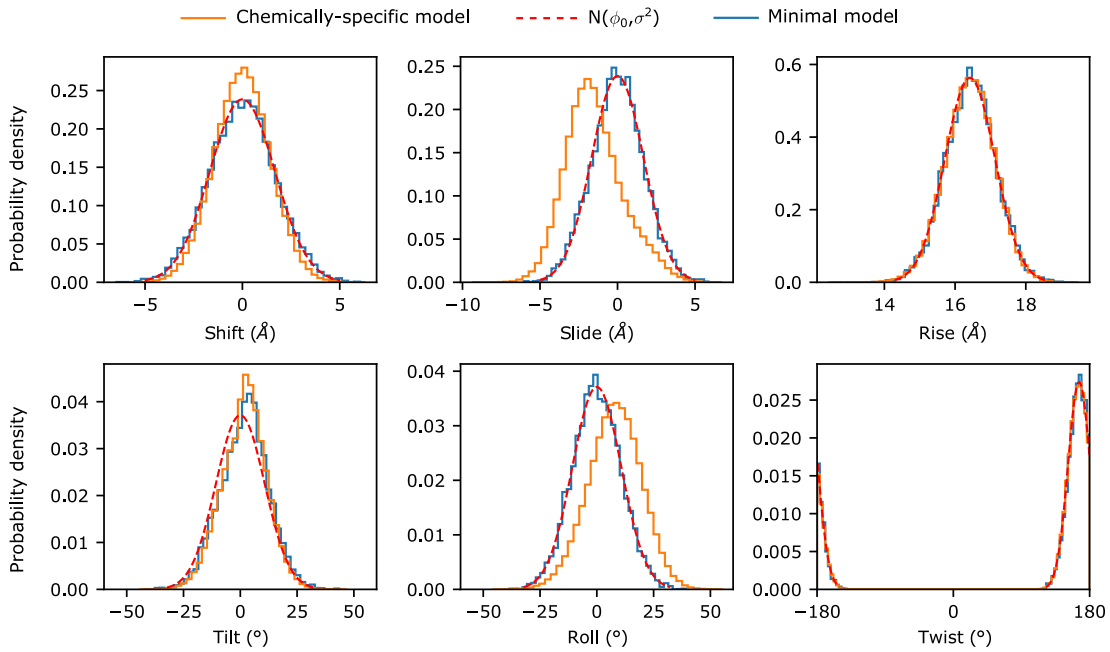

Figure 5.6: **Helical parameter distributions between minimal model DNA beads (1 bead per 5 base-pairs) from chemically-specific model simulations (orange line) and minimal model simulations (blue line)**. The red dashed lines are normal distributions using the mean and variance of each helical parameter.

between the minimal model DNA beads were computed, the distributions are plotted as the orange lines in 5.6.

We found that twist has a mean of 167 degrees which results in the distribution wrapping around from 180 to -180, this introduces some computational problems that do not occur for the standard rigid base pair model where the twist mean value is ∼30 and is too stiff to ever wrap around past 180. The first issue is the potential energy calculation in equation 5.9 computes $\Delta\phi_{\text{twist}}$ as simply $\phi_{\text{twist}} - \phi_{0,\text{twist}}$, clearly when ϕ_{twist} crosses from 180 to -180 this no longer correctly computes the difference in angles that is relevant for the potential, that is we always want the smaller angle and never 360 minus the angle. To resolve this issue we always compute angular differences using the function *angleDiff(a,b)* which computes $b - a$ constrained to the range $[-180, 180)$.

```
angleDiff(a,b):
    d =  fmod(b-a+180.0, 360.0) # floating point remainder
    if d < 0.0:
        d += 360.0
    return d - 180.0
```

The second issue is that when the value of twist transitions from 180 to -180 the values of shift, slide, tilt and roll, as calculated by the procedure in section A.2, all flip sign. Once again due to the definition of the potential this can cause discontinuities. However, we observe that the potential is continuous under helical parameter sign flip if the equilibrium values are 0 and **K** is diagonal. Fortunately, looking at figure 5.6 we see that the distributions of shift, slide, tilt, and roll are approximately centered around zero, therefore we make this approximation and set them to 0. For rise and

twist we set the equilibrium values to the means of the sampled distributions.

$$\boldsymbol{\phi}_0 = (0, 0, 16.44, 0, 0, 166.69). \tag{5.14}$$

Additionally, we set $\mathbf{K}$ to be diagonal by simply ignoring the off diagonal terms and setting them to zero, this approximation is appropriate as the off-diagonal terms are an order of magnitude less than the diagonals (Clauvelin et al [82] make a similar approximating when using the standard RBP). The computed value of $\mathbf{K}$ is

$$\mathbf{K} = \text{diag}(0.301, 0.235, 1.56, 0.00614, 0.00515, 0.00724). \tag{5.15}$$

The probability density functions for each parameter are normal distributions, $N(\phi_{0i}, C_{ii})$, plotted as the red curves in figure 5.6. Finally as a sanity check, we ran the minimal model and computed its helical parameter distributions, plotted as the blue curve in figure 5.6.

5.5.2 Pairwise terms

5.5.2.1 LJ interaction

The main term of the pairwise interactions is a shifted and truncated Lennard-Jones term, this accounts for both excluded volume and attractive forces.

$$E_{\text{Minimal-LJ}} = \begin{cases} E_{\text{LJ}}(r) - E_{\text{LJ}}(r_c^{\text{LJ}}), & r \leq r_c^{\text{LJ}}, \\ 0, & r > r_c^{\text{LJ}}, \end{cases} \tag{5.16}$$

$$E_{\text{LJ}}(r) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right], \tag{5.17}$$

where ϵ is the interaction strength, σ is the zero crossing point, r is the distance between the pair of interacting particles, and r_c^{LJ} is the cutoff distance. We parameterize this potential to represent the DNA to histone core electrostatic attraction and the DNA to DNA electrostatic repulsion that is present in the chemically-specific model. To account for the salt dependent behavior of chromatin we fit empirical relationships between the simulated salt concentration c and the values of ϵ and σ in this potential.

5.5.2.2 Fitting parameters of the LJ interaction

To fit the parameters of the minimal LJ model we aim to reproduce two key features of the chemically-specific model simulations. The first is the orientation dependent nucleosome-nucleosome interactions, the second is the salt dependent behavior of the radius of gyration of a 12-nucleosome chromatin fiber.

We first computed the inter-nucleosome PMFs for the chemically-specific model for high and low salt, these are shown in figure 5.7 a1 and a2 (these are the same results as in section 4.8). We then performed a similar calculation with the minimal model. Due to the fact that the minimal nucleosomes are completely rigid, umbrella sampling is not need, instead we can do a direct calculation of the potential energy as a function of the inter-nucleosome distance. We optimize the values of σ and ϵ such that the minimal model interaction curves best approximate the shape of chemically-specific model curves. With these initial guesses of σ and ϵ, which are at the end

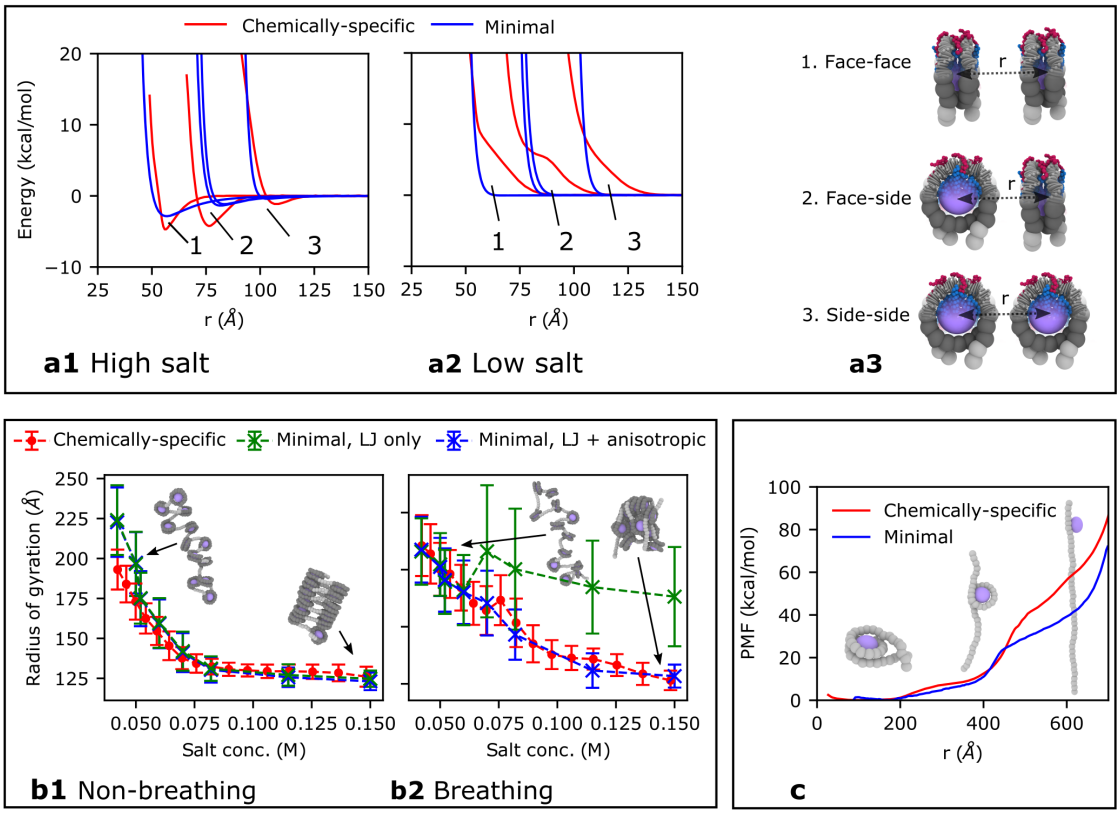

Figure 5.7: **Fitting minimal model pairwise terms to the chemically-specific model**. **(a)** Shows the inter-nucleosome PMFs computed with the chemically-specific and minimal models for high (a1) and low (a2) salt. The 3 orientations are illustrated in (a3) with the collective variable r labelled. The chemically-specific model curves are used to set the σ values in the minimal model by manually adjusting the values until the minima and diverging repulsive section of the curves have good agreement. **(b)** Shows salt dependent radius of gyration for the chemically-specific model, the minimal model with just the LJ interaction, and with the LJ plus anisotropic interactions. The plots are for the non-breathing (b1) and breathing (b2) nucleosome configurations. The non-breathing curves are used to fit the ϵ by performing grid searches to give the minimal curves that best match the chemically-specific curve. (b2) Demonstrates that the anisotropic potential is needed to recover the behavior of breathing chromatin. **(c)** shows the PMF of nucleosome unwrapping for the chemically-specific and minimal model nucleosomes. The values of the energy in the anisotropic potential were optimized by attempting to best match the PMF curves in the low extension regime (< 450Å).

points of the salt range, we proceed to find an adequate interpolation to model the salt dependent behavior. To do this we use the radius of gyration of non-breathing 12-nucleosome chromatin as the observable to compare between the minimal model and the chemically-specific model. The chemically-specific model radii of gyration come from the simulations in section 4.9. To enable reliable comparison the radius of gyration is computed from just the center of mass coordinates of the nucleosomes. Using a combination of manual adjustment and grid search techniques we obtained the optimal parameters (in table 5.2). These parameters give the R_g values in figure 5.7b2 which compare well with the chemically-specific model.

Moving on to the breathing model and comparing the radii of gyration with the chemically-specific model (red and green lines in figure 5.7b2), we find a significant difference in the behavior at higher salt. This is due to the unwrapped DNA not having strong enough interactions to the exposed nucleosome cores. To account for this we developed the anisotropic potential which provides a short range attractive potential mimicking the DNA binding region.

Interaction pair	ϵ (kcal/mol)	σ (Å)	r_c^{LJ} (Å)
Core–core	0.1	55	$2^{1/6}\sigma$
Core–DNA	$E(c)$	40	3σ
DNA–DNA	0.1	$S(c)$	$2^{1/6}\sigma$

Table 5.2: Minimal model Lennard-Jones parameters. c is the salt concentration. $E(c)$ and $S(c)$ are linear interpolations of the data in table 5.3. Note that $r_c^{\mathrm{LJ}} = 2^{1/6}\sigma$ implies that the potentials are repulsive only.

c (mol/L)	$E(c)$ (kcal/mol)	$S(c)$ (Å)
0.15	0.4	24
0.115	0.375	24.5
0.082	0.35	25
0.07	0.3	26
0.06	0.25	27
0.052	0.2	28
0.05	0.1	30
0.042	0.01	34

Table 5.3: $E(c)$ and $S(c)$ are found by linear interpolation of this data.

5.5.2.3 Anisotropic interaction

The anisotropic term is a pairwise potential that depends on the relative orientation and shape of the interacting pair of ellipsoidal particles:

$$E_{\text{anisotropic}} = U_r(\mathbf{A}_i, \mathbf{A}_j, \mathbf{r})\eta(\mathbf{A}_i, \mathbf{A}_j, \mathbf{r})\chi(\mathbf{A}_i, \mathbf{A}_j, \mathbf{r}), \tag{5.18}$$

where $\mathbf{A}_i$ and $\mathbf{A}_j$ are the orientation matrices of particles i and j with center to center separation vector $\mathbf{r}$. This potential is a modified version of the well-known

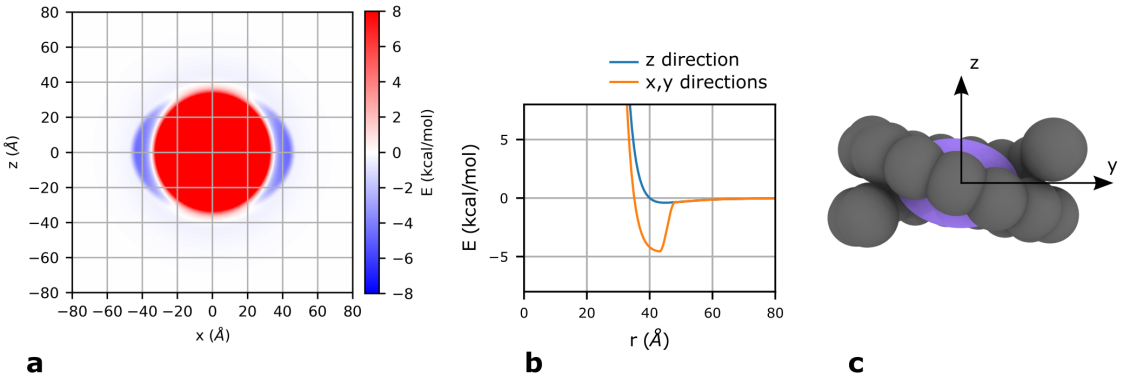

Figure 5.8: **Minimal pairwise interactions**. (a,b) Show the resulting potential felt by the DNA beads due to the core beads. The attractive region is only in the x-y plane around the core where the DNA binds in a typical nucleosome. (c) Shows an illustration of the nucleosome in the same orientation as the plot in a.

Gay-Berne potential [195] where we have replaced the Lennard-Jones like term with a cosine-squared term [196]. This allows for greater control over the depth and range of the potential.

We include the anisotopic interaction to account for DNA binding to the exposed nucleosome core in the breathing nucleosome simulations. Our radii of gyration plotted in figure 5.7b2 demonstrate the need for this, at high salt breathing chromatin has a much higher radius of gyration than desired. This is because breathing chromatin has significantly more DNA beads classed as linker DNA, these are not rigidly fixed to the core, and the pair wise LJ terms are not strong enough to account for the binding of the unwrapped linker DNA to the exposed nucleosome cores. To model the DNA binding we desire a potential that is only attractive in the regions where the DNA binds to the core. We know that this region is around the sides of the nucleosome core and not on the faces where the acid patch is located, thus the anisotropic potential enables this. Figure 5.8 demonstrates the resulting potential due to this term.

The U_r term controls the interaction based on the distance of closest approach between the ellipsoids h, in the standard GB potential this term has the form of a LJ interaction, in ours it is:

$$U_r = \begin{cases} -\epsilon & h < 0, \\ -\epsilon \left(\cos\left(\frac{\pi h}{2 r_c^{\text{aniso}}} \right) \right)^2 & 0 \leq h < r_c^{\text{aniso}}, \\ 0 & h \geq r_c^{\text{aniso}}, \end{cases} \tag{5.19}$$

where h the distance of closest approach is given by

$$h = r - \sigma, \tag{5.20}$$

$$\sigma = \left[\frac{1}{2} \hat{\mathbf{r}}^{\intercal} \mathbf{G}^{-1} \hat{\mathbf{r}} \right]^{-1/2}, \tag{5.21}$$

$$\mathbf{G} = \mathbf{G}_i + \mathbf{G}_j = \mathbf{A}_i^{\intercal} \mathbf{S}_i^2 \mathbf{A}_i + \mathbf{A}_j^{\intercal} \mathbf{S}_j^2 \mathbf{A}_j, \tag{5.22}$$

where $\mathbf{S}_i = \text{diag}(a_i, b_i, c_i)$ is the shape matrix of particle i given by the ellipsoid radii. The η and χ terms are dimensionless scaling terms that depend on the relative orientation, shape, and form of each particles relative energy matrix. They are unchanged from the versions in LAMMPS [197].

$$\eta = \left[\frac{2s_i s_j}{det(\mathbf{G})}\right]^{\nu/2}, \tag{5.23}$$

$$s_i = [a_i b_i + c_i c_i][a_i b_i]^{1/2}, \tag{5.24}$$

$$\chi = \left[\hat{\mathbf{r}}^\intercal \mathbf{B}^{-1} \hat{\mathbf{r}}\right], \tag{5.25}$$

$$\mathbf{B} = \mathbf{B}_i + \mathbf{B}_j = \mathbf{A}_i^\intercal \mathbf{E}_i \mathbf{A}_i + \mathbf{A}_j^\intercal \mathbf{E}_j \mathbf{A}_j, \tag{5.26}$$

where $\mathbf{E}_i = \text{diag}(1/\epsilon_{ai}, 1/\epsilon_{bi}, 1/\epsilon_{ci})$ is the relative energy matrix of ellipsoid i given by the inverse well depths.

Parameter	Value
$\mathbf{S}_{\text{Core}}$	$\text{diag}(28, 28, 20)$
$\mathbf{S}_{\text{DNA}}$	$\text{diag}(12, 12, 12)$
ϵ	6 kcal/mol
r_c	5 Å
$\mathbf{E}_{\text{core}}$	$\text{diag}(1, 1, 1/0.0001)$
$\mathbf{E}_{\text{DNA}}$	$\text{diag}(1, 1, 1)$

Table 5.4: Anisotropic potential parameters. diag(a,b,c) means diagonal a 3x3 matrix with the elements a,b,c on the diagonal.

5.5.2.4 Forces and Torques

The forces and torques for the anisotropic potential are a slightly modified version of the Gay-Berne forces and torques in LAMMPS [197, 198], the only differences are fixing ν and μ as 1, and the form of U_r. For completeness we will give an overview of the derivations and the state the results, for full details [197, 198] should be consulted.

The force is given by:

$$\mathbf{f} = -\frac{\partial E_{\text{Anisotropic}}}{\partial \mathbf{r}} = -\eta \left(U_r \frac{\partial \chi}{\partial \mathbf{r}} + \chi \frac{U_r}{\partial \mathbf{r}}\right). \tag{5.27}$$

The derivatives can be split into components parallel to and perpendicular to the inter-particle displacement vector $\mathbf{r}$, for a general pair potential this has the form:

$$\frac{\partial U}{\partial \mathbf{r}} = \frac{\partial U}{\partial r}\hat{\mathbf{r}} + r^{-1}\frac{U}{\partial \hat{\mathbf{r}}} \cdot (\mathbf{1} - \hat{\mathbf{r}} \otimes \hat{\mathbf{r}}), \tag{5.28}$$

where $\hat{\mathbf{r}} \otimes \hat{\mathbf{r}}$ denotes the outer product of the two vectors. For the U_r term the variables

$$\mathbf{k} = \mathbf{G}^{-1}\mathbf{r}, \tag{5.29}$$

and

$$\varphi = \sigma^{-2} = \frac{1}{2}\hat{\mathbf{r}}^\intercal \mathbf{G}^{-1} \hat{\mathbf{r}}, \tag{5.30}$$

are introduced. This gives

$$\frac{\partial U_r}{\partial \mathbf{r}} = \frac{\partial U_r}{\partial r}\hat{\mathbf{r}} + r^{-2}\frac{\partial U_r}{\partial \varphi}\left[\mathbf{k} - (\mathbf{k}\cdot\hat{\mathbf{r}})\hat{\mathbf{r}}\right], \tag{5.31}$$

where

$$\frac{\partial U_r}{\partial \varphi} = \frac{\sigma^3}{2}\frac{U_r}{\partial r}, \tag{5.32}$$

$$\frac{\partial U_r}{\partial r} = \begin{cases} 0, & h < 0, \\ \frac{\pi\epsilon}{2r_c}\sin\left(\frac{\pi h}{r_c}\right) & 0 \le h < r_c^{\text{aniso}}, \\ 0, & h \ge r_c^{\text{aniso}}. \end{cases} \tag{5.33}$$

Similarly for the χ term

$$\frac{\partial \chi}{\partial \mathbf{r}} = -4r^{-2}\left[\mathbf{l} - (\mathbf{l}\cdot\hat{\mathbf{r}})\hat{\mathbf{r}}\right], \tag{5.34}$$

where

$$\mathbf{l} = \mathbf{B}^{-1}\mathbf{r}. \tag{5.35}$$

The torque can be computed by first considering the derivative of a general pair potential U with respect to rotating the particle about an axis $\hat{\boldsymbol{\psi}}$:

$$\hat{\boldsymbol{\psi}}\cdot\boldsymbol{\tau} = -\frac{\partial U}{\partial \psi} = -\sum_m \frac{\partial U}{\partial \hat{\mathbf{a}}_m}\cdot\frac{\partial \hat{\mathbf{a}}_m}{\partial \psi} = \sum_m \frac{\partial U}{\partial \hat{\mathbf{a}}_m}\cdot\hat{\boldsymbol{\psi}}\times\hat{\mathbf{a}}_m = \hat{\boldsymbol{\psi}}\cdot\sum_m \hat{\mathbf{a}}_m\times\frac{\partial U}{\partial \hat{\mathbf{a}}_m}, \tag{5.36}$$

where $\hat{\mathbf{a}}_m$ are the axis unit vectors (columns) of the particles orientation matrix $\mathbf{A}$. Setting $\hat{\boldsymbol{\psi}}$ to be the x, y, and z axis gives

$$\boldsymbol{\tau} = -\frac{\partial U}{\partial \mathbf{q}} = -\sum_m \hat{\mathbf{a}}_m\times\frac{\partial U}{\partial \hat{\mathbf{a}}_m}. \tag{5.37}$$

The torque from the anisotropic potential on particle i is

$$\boldsymbol{\tau}_i = -\eta\chi\frac{\partial U_r}{\partial \mathbf{q}_i} - U_r\chi\frac{\partial \eta}{\partial \mathbf{q}_i} - U_r\chi\frac{\partial \chi}{\partial \mathbf{q}_i}. \tag{5.38}$$

where

$$\frac{\partial U_r}{\partial \mathbf{q_i}} = -r^{-2}\frac{\partial U_r}{\partial \varphi}\left(\boldsymbol{\kappa}^\intercal\mathbf{G}_i\right)\times\boldsymbol{\kappa}, \tag{5.39}$$

$$\frac{\partial \chi}{\partial \mathbf{q}_i} = 4r^{-2}(\mathbf{l}^\intercal\mathbf{B}_i)\times\mathbf{l}, \tag{5.40}$$

$$\frac{\partial \eta}{\partial \mathbf{q}_i} = -\sum_m \mathbf{a}_{im}\times\mathbf{d}_{im}, \tag{5.41}$$

where $\mathbf{d}_{im}$ is column m of matrix $\mathbf{D}_i$,

$$\mathbf{D}_i = -\frac{1}{2|\mathbf{G}|}\left(\frac{2s_1s_2}{|\mathbf{G}|}\right)^{1/2}\mathbf{E}_i, \tag{5.42}$$

where

$$\mathbf{E}_i = \frac{\partial|\mathbf{G}|}{\partial \mathbf{A}_i}, \tag{5.43}$$

which has components [197]

$$E_{jk} = |\mathbf{G}|\text{Trace}\left[\mathbf{G}^{-1}(\hat{\mathbf{e}}_j \otimes \hat{\mathbf{a}}_k + \hat{\mathbf{a}}_k \otimes \hat{\mathbf{e}}_j)S_{kk}^2\right], \tag{5.44}$$

where $\hat{\mathbf{e}}_j$ are the unit axis vectors of the lab frame ($\hat{\mathbf{e}}_1 = (1,0,0), \hat{\mathbf{e}}_2 = (0,1,0), \hat{\mathbf{e}}_3 = (0,0,1)$). This last expression uses the relation for determinant derivatives:

$$\partial|\mathbf{G}| = |\mathbf{G}|\text{Trace}\left[\mathbf{G}^{-1}\partial\mathbf{G}\right], \tag{5.45}$$

and $\partial\mathbf{G}/\partial\mathbf{A}$ is implied by the definition of $\mathbf{G}$ in terms of $\mathbf{A}$ in equation 5.22 [198].

5.5.2.5 Fitting parameters of the anisotropic interaction

To fit the parameters of the anisotropic potential we first use our knowledge of the shape of the nucleosome core and the region where the DNA binds. This allows us to set the values of the ellipsoid shape matrices $\mathbf{S}_{\text{core}}$ and $\mathbf{S}_{\text{DNA}}$, and the energy matrices $\mathbf{E}_{\text{core}}$ and $\mathbf{E}_{\text{DNA}}$. Specifically $\mathbf{E}_{\text{core}}$ is constructed such that there is no attraction at the z-poles of the ellipsoid by setting $\epsilon_{\text{core},z}$ to 0.00001 and all the other relative well depths to 1. We then ensure that the interaction is short range enough to have no effect on the non-breathing model, this means the potential well is completely covered by the bound nucleosomal DNA. This gives us r_c^{aniso}; the correspondence of the green and blue points in figure 5.7b1 demonstrates the value ensures the anisotropic potential has no effect on the non-breathing minimal model. To fit the remaining parameter ϵ_{aniso}, the depth of the potential, we computed the PMF of nucleosome unwrapping for a minimal model nucleosome where the DNA is completely free from the core and only the pairwise interactions keep it bound. We compare this with the nucleosome unwrapping PMF computed using the chemically-specific model (detailed in section 4.5) and, using a grid search parameter sweep, we find the value of 6 kcal/mol gives the best agreement in the low extension regime. The PMFs are shown in figure 5.7c.

5.6 Computational implementation in LAMMPS

We have implemented the anisotropic potential in the source code files `pair_aniso.cpp` and `pair_aniso.h` included in our code repository, they are both modified from the existing LAMMPS code `pair_gayberne.cpp` and `pair_gayberne.h` from the `ASPHERE` package. The minimal-RBP potential uses the same code as the chemically-specific model, the `NAFlex_params.txt` file just needs to be be changed to contain the minimal-RBP parameters.

The LAMMPS input scripts are very similar to the chemically-specific model:

```
units real
atom_style hybrid_ellipsoid angle charge
```

The bonds style is

```
bond_style harmonic/DNA
bond_coeff 1 0 0
```

which reads in the text files `NAFlex_params.txt` and `DNA_seqeunce.txt`.

The pair style is

```
pair_style hybrid/overlay lj/cut LJCUT ansio 1 1 1 ANISOCUT
```

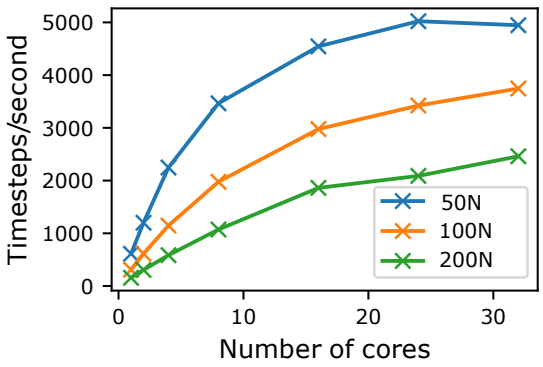

Figure 5.9: Performance of the minimal model for 50, 100, 200 nucleosome chromatin in timesteps per second.

this combines a standard LJ potential and our ansiotropic potential, the parameters LJCUT and ANISOCUT are the overall cutoff distances used in the potentials. For `lj/cut` this will be the largest value out of the specific pairwise cutoffs and for `aniso` we set it to 60Å. The pair coefficients are set as follows

`pair_coeff 1 1 lj/cut` $\epsilon_{\text{core-core}}$ $\sigma_{\text{core-core}}$ $r^{\text{LJ}}_{\text{c, core-core}}$
`pair_coeff 1 2 lj/cut` $\epsilon_{\text{core-dna}}$ $\sigma_{\text{core-dna}}$ $r^{\text{LJ}}_{\text{c, core-dna}}$
`pair_coeff 1 3 lj/cut` $\epsilon_{\text{core-dna}}$ $\sigma_{\text{core-dna}}$ $r^{\text{LJ}}_{\text{c, core-dna}}$
`pair_coeff 2 2 lj/cut` $\epsilon_{\text{dna-dna}}$ $\sigma_{\text{dna-dna}}$ $r^{\text{LJ}}_{\text{c, dna-dna}}$
`pair_coeff 2 3 lj/cut` $\epsilon_{\text{dna-dna}}$ $\sigma_{\text{dna-dna}}$ $r^{\text{LJ}}_{\text{c, dna-dna}}$
`pair_coeff 3 3 lj/cut` $\epsilon_{\text{dna-dna}}$ $\sigma_{\text{dna-dna}}$ $r^{\text{LJ}}_{\text{c, dna-dna}}$
`pair_coeff 1 2 aniso` ϵ_{ansio} `0` $\epsilon_{\text{core,x}}$ $\epsilon_{\text{core,y}}$ $\epsilon_{\text{core,z}}$ $\epsilon_{\text{dna,x}}$ `\`
$\epsilon_{\text{dna,y}}$ $\epsilon_{\text{dna,z}}$ r_c^{ansio} `ANISOCUT`
`pair_coeff 1 2 aniso` ϵ_{ansio} `0` $\epsilon_{\text{core,x}}$ $\epsilon_{\text{core,y}}$ $\epsilon_{\text{core,z}}$ $\epsilon_{\text{dna,x}}$ `\`
$\epsilon_{\text{dna,y}}$ $\epsilon_{\text{dna,z}}$ r_c^{ansio} `ANISOCUT`

Where for clarity we have explicitly listed all combinations. The particle type numbers are in table 5.1 and the values of the parameters are in tables 5.2 and 5.4. The neighbour list skin distance is set as

```
neighbour 50 bin
```

The integration settings are

```
comm_style tiled
fix bl all balance 1000 1.0 rcb
fix 1 nucl rigid/nve/small molecule
fix 2 linker_dna  nve/asphere
fix 3 linker_dna  langevin 300 300 5000000 123 angmom 3.0
fix 4 nucl  langevin 300 300 5000000 123
timestep 500
run N
```

where the group `nucl` contains the core and nucleosomal DNA particles. The difference is that the `nve/apshere` integrator is used for the linker DNA beads as they are single ellipsoids. The timestep has been set to 500 fs which is 5 times the timestep we found appropriate for the DNA model in section 4.1.1.

The performance of the minimal model for different system sizes is shown in figure 5.9.

Chapter 6

Minimal-model simulations

In this chapter we report the results from simulations performed using the minimal model. Part of these results are in [121].

Contents

6.1 Timescales

6.1.1 Diffusion coefficients

To assess the timescale difference between our two models and the timescale differences between our models and experiments we computed the diffusion coefficient of DNA as a function of length. This is the diffusion coefficient of DNA in water in the

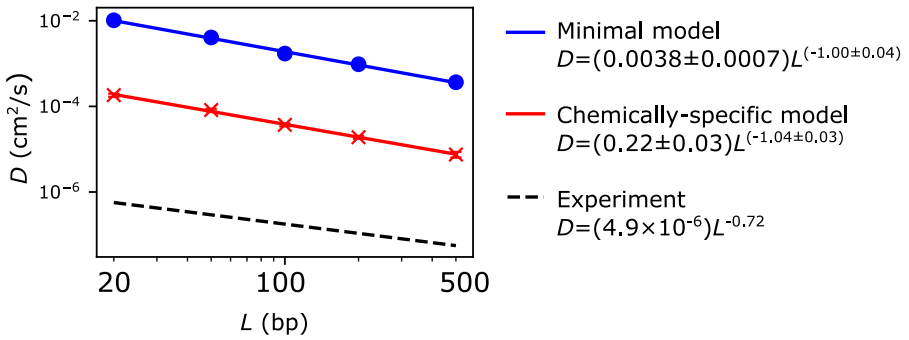

Figure 6.1: Diffusion coefficients D of DNA as a function of DNA polymer length L. Computed using the minimal model (blue line), chemically-specific model (red line), and an empirical fit to experimental data taken from literature [199].

limit of infinite dilution, i.e the DNA molecule only interacts with the solvent, not other DNA molecules. To compute the diffusion coefficient we did multiple simulations (64 repeats) of a single DNA strand in the NVT ensemble and measured the mean squared displacement (MSD)

$$\mathrm{MSD}(t) = \langle |\mathbf{r}(t) - \mathbf{r}(0)|^2 \rangle, \tag{6.1}$$

where $\mathbf{r}(t)$ is the position of the center of mass of the DNA strand at time t with initial position $\mathbf{r}(0)$. The average is taken over our 64 repeat simulations. Plotting the MSD against time and fitting a straight line gives the diffusion coefficient D as one sixth of the gradient

$$\mathrm{MSD}(t) = 6Dt. \tag{6.2}$$

We computed D for the chemically-specific model and the minimal model for DNA lengths of 20, 50, 100, 200, and 500 bp. The values are plotted in figure 6.1, along with an empirical fit of D from measurements of DNA in water [199], where power law equations have been fitted to asses the relationship between D and the length of the DNA L

$$D \propto L^{\nu}. \tag{6.3}$$

Both our values of ν are ≈ -1, the value from experiment is -0.72. The discrepancy could be explained by our smaller DNA lengths of 20-500bp vs 21-6000bp, additionally, due to our Langevin dynamics simulation method, by measuring D we are measuring what we put into the simulation. The diffusion coefficient is given by the the Einstein relation

$$D = \frac{k_B T}{\gamma}, \tag{6.4}$$

where γ is the friction coefficient that also appears in the Langevin equation (equation 2.20), the value of $\nu = -1$ implies the friction $\propto L$, this makes conceptual sense, doubling the number of beads doubles the friction.

Comparing the values of D for the same L gives the approximate relations

$$D_{\text{Minimal}} \sim 50 D_{\text{Chemically-specific}}, \; D_{\text{Chemically-specific}} \sim 100 D_{\text{Experiment}}. \tag{6.5}$$

This implies 1 ns in the chemically-specific model represents 100 ns in "real" time and 1 ns in the minimal model is equivalent to 50 ns in the chemically-specific model.

6.1.2 Autocorrelation of radius of gyration

As an alternate assessment of the timescale difference between the models we computed the autocorrelation function of the radius of gyration of a 12-nucleosome chromatin system at low salt conditions in both models. The radius of gyration is calculated from the center of masses of the nucleosomes. The time series are plotted in 6.2a and b for the chemically-specific model and the minimal model respectively. We computed the autocorrelation function $C(\tau)$ using

$$C(\tau) = \frac{\langle (R_g(t) - \langle R_g \rangle)\,(R_g(t+\tau) - \langle R_g \rangle) \rangle}{\mathrm{Var}(R_g)}, \tag{6.6}$$

where τ is the time lag, $R_g(t)$ are the values of the radius of gyration at timestep t, $\langle R_g \rangle$ is the mean radius of gyration, $\mathrm{Var}(R_g)$ is the variance, and the averages

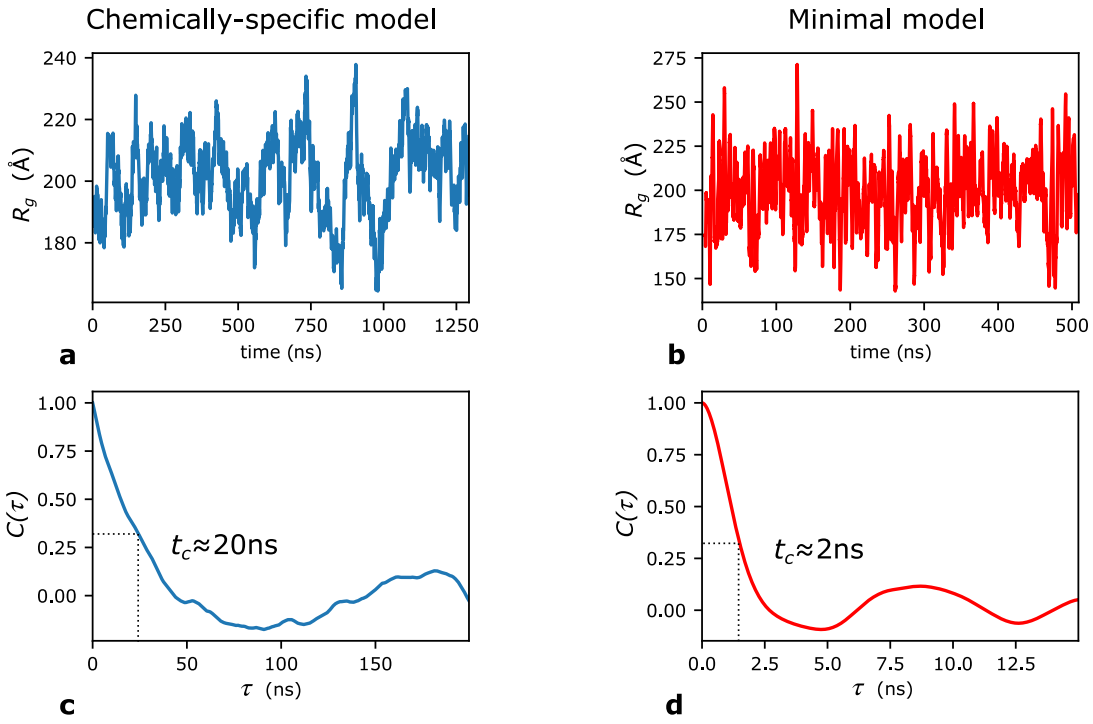

Figure 6.2: **Timescale comparison between chemically-specific and minimal models. (a-b)** Timeseries of the radius of gyration R_g for 12-nucleosome chromatin using the chemically-specific model and the minimal model respectively. **(c-d)** Autocorrelation functions $C(\tau)$ of R_g for the chemically-specific model and minimal model respectively. The correlation time t_c is labelled, it is estimated from where the curves reach a height of $1/e$.

are taken over t. The auto-correlation functions are plotted in 6.2 c and d, the values of the correlation time t_c are approximated by reading off the graphs where the value of $C(\tau)$ reaches $1/e$. We find that the correlation time for the chemically-specific model is 20 ns, while for the minimal model it is 2 ns, this suggest that timescales in the minimal model are 10 times faster than in the chemically-specific model. Note that this is different to the factor of 50 we found in the previous section and is not unexpected, different dynamics have different timescales, and the changes in these timescales as we coarse-grain have no reason to be equal changes — Langevin dynamics does not "correctly" simulate dynamics, it only correctly simulates equilibrium properties in the canonical ensemble.

6.2 Impact of nucleosome breathing on liquid-liquid phase separation of chromatin

Recent experiments have discovered that 12-nucleosome reconstituted chromatin arrays undergo intrinsic LLPS (i.e without the aid of additional proteins) under physiological salt concentrations in vitro and when microinjected into cells [31]. Extensive studies characterizing LLPS of proteins and nucleic acids have demonstrated that multivalency is the dominant driving force for their LLPS [46, 200–204]. That is proteins with high valencies (where valency is defined as the number of other molecules a molecule simultaneously interacts with) can stabilize LLPS by form-

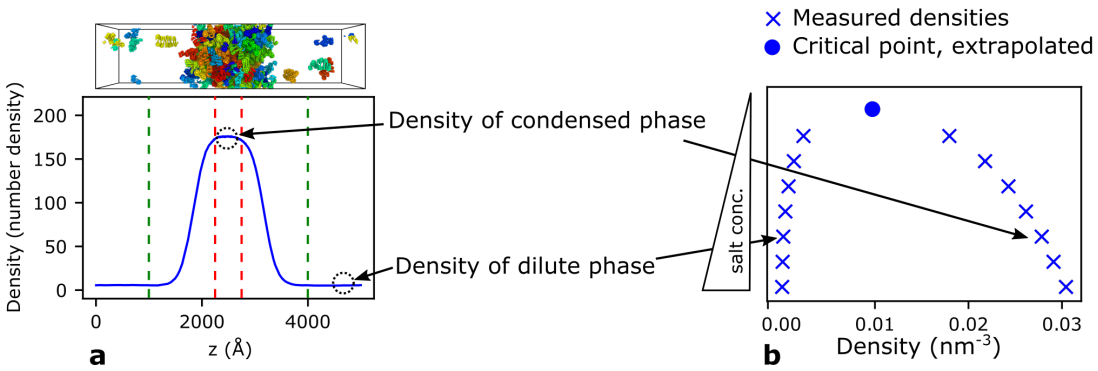

Figure 6.3: **Method for computing a LLPS phase-diagram from coexistence simulations.** **(a)** A density profile computed from a coexistence simulation, an example simulation snapshot is shown above the plot. The values within the dashed red lines are used to compute the condensed phase density, the values outside the green dashed lines are used to compute the density of the dilute phase. One density profile gives a pair of phase-diagram data points. **(b)** A liquid-liquid phase diagram, each pair of data points at a specific salt is computed from a coexistence simulation. The critical point is estimated using equations 6.7 and 6.8.

ing numerous protein–protein [200, 202, 203], protein–RNA [205], and/or protein–DNA [28, 29] interactions, that compensate for the entropy loss upon demixing [204]. Furthermore, the binding of multiple Swi6 proteins to nucleosomes was shown to facilitate LLPS of chromatin [40]. The mechanism behind this involves the Swi6 reshaping the nucleosomes, exposing the histone cores, in a manner consistent with nucleosome breathing. These ideas, together with the valency enhancement we see from our simulations of breathing chromatin, led us to hypothesize that nucleosome breathing is important for facilitating the intrinsic LLPS of chromatin.

To investigate this phenomenon, we deployed our minimal coarse-grained chromatin model as it is capable of simulating the system sizes required (hundreds of chromatin arrays) to capture the collective phase behavior of chromatin. Specifically, we performed direct coexistence simulations of systems containing 125 independent 12-nucleosome chromatin arrays with a uniform 165 NRL at different monovalent salt conditions for both breathing and non-breathing chromatin. From these simulations we computed the liquid-liquid phase diagrams of breathing and non-breathing chromatin at 300K in 'NaCl concentration' versus 'chromatin density' space.

6.2.1 Methods

In order to compute the phase diagram of 12-nucleosome chromatin arrays, we employ the direct coexistence method [206–209] using 125 independent 12-nucleosome chromatin arrays with 165-bp NRL at different monovalent salt conditions. The initial structures for the chromatin arrays are obtained by generating them from randomly selected, equilibrated, structures from the chemically-specific model HREMD simulations at the corresponding salt (version 1 in section 5.4).

In a direct coexistence simulation, illustrated in figure 6.3, one places both phases; i.e., the dilute liquid and condensed liquid phase in the same simulation box. The simulation is performed until they reach equilibrium at their coexistence

densities. Once equilibrium is reached, the densities of coexistence are determined by computing the average density profile along the z axis of the simulation box.

The average density profiles are computed by first constructing a density histogram, along the simulation box z direction, for each timestep in the trajectory. We use a bin width of 5Å. The histograms are centered on the systems center of mass at the corresponding time-step. We then average the histograms over all timesteps.

We estimate the critical salt concentration c_c, by fitting the density difference between the coexisting low-density $\rho_l(c)$ and high-density $\rho_h(c)$ phases to the expression [210],

$$\left(\rho_h(c) - \rho_l(c)\right)^{3.06} = d\left(1 - \frac{c}{c_c}\right), \tag{6.7}$$

where d is a fitting parameter. The critical density ρ_c is estimated using the law of rectilinear diameter,

$$\left(\rho_l(c) + \rho_h(c)\right)/2 = \rho_c + s\left(c_c - c\right), \tag{6.8}$$

where s is a fitting parameter.

To prepare the initial configurations for the systems we used the slab method [133]. This involves setting up the particles in a periodic simulation box that is large enough such that there is no overlap between molecules. First in the NVT ensemble the x and y box dimensions are slowly scaled to the target size of 1200Å. The simulation is then run in the NPT ensemble for 100 ns where the z dimension is coupled to a Berendsen barostat [211] set to a high enough pressure (typically 1 bar) to compress the system. Once compressed the z dimension of the periodic box is then increased to 5000Å. We then conducted the production NVT simulations, in a periodic box of size 1200Å $\times$ 1200Å $\times$ 5000 Å, for several µswith coordinate snapshots were recorded every 10,000 timesteps. The last half of the trajectories were used for analysis.

The density profiles from our simulations are shown in figure 6.4. The values of density for the high and low density phases were calculated as the means of the data points between the dashes lines. The standard deviations were also calculated and are indicated on the phase-diagrams by the error bars which are too small to be visible for most densities.

6.2.1.1 Estimation of liquid-network connectivity

We define the connectivity of the liquid-network as the mean number of connections per chromatin array in the high-density phase multiplied by the density of the high-density phase, which gives the number of inter-chromatin bonds per unit volume. The number of connections of a chromatin array is defined as the number of distinct chromatin arrays it is in contact with. Two nucleosomes are defined to be "in contact" if the nucleosome–nucleosome distance (i.e. any nucleosome in one chromatin array relative to any nucleosome in another chromatin array) is less than 110 Å.

6.2.2 Results and Discussion

Our computed phase diagrams of chromatin LLPS are plotted in figure 6.5a and reveal that the enhancement of nucleosome valency due to nucleosome breathing

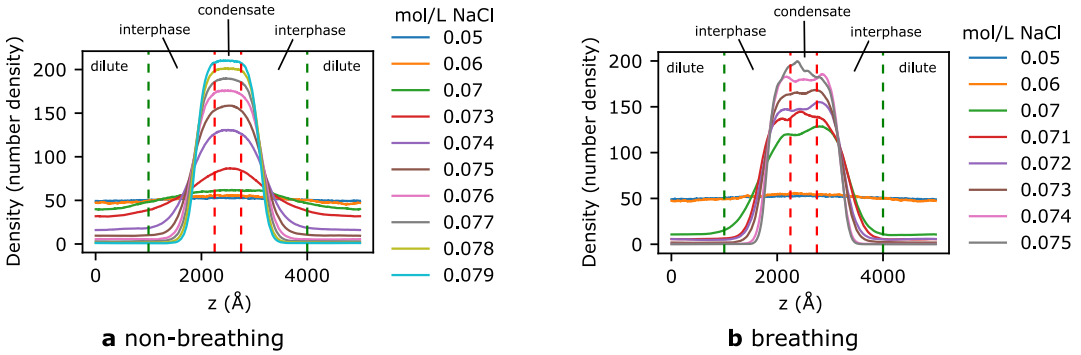

Figure 6.4: **Density profiles of direct coexistence simulations of 12-nucleosome chromatin**. The density profiles at a range of salt values are plotted for non-breathing **(a)** and breathing chromatin **(b)**. The density of the condensed phase is computed from the values between the red dashed lines, the density of the dilute phase is computed from the values outside the green dashed lines.

increases the range of stability of the intrinsic LLPS of chromatin. That is, chromatin forms condensates above a critical monovalent salt concentration, due to the screening by counterions becoming sufficiently strong to eliminate the DNA–DNA repulsion and enabling the formation of numerous attractive inter-nucleosome interactions. Compared with breathing nucleosomes, constraining the nucleosomes to be non-breathing, results in chromatin requiring higher NaCl concentration for LLPS to become thermodynamically stable; the limited valency of the non-breathing chromatin arrays, requires more salt to sufficiently screen the DNA–DNA repulsion and allow the inter-fiber nucleosome interactions to facilitate LLPS. We believe that this occurs because in the non-breathing chromatin, nucleosomes are perfectly stacked face-to-face, and intra-fiber interactions are favored over inter-fiber ones. At the same salt concentration, breathing chromatin produces a more dense, more stable condensed liquid. Snapshots of breathing and non-breathing coexistence simulations at the same salt concentration are shown in figure 6.6. For further analysis of chromatin LLPS we computed the connectivity (bonds per unit volume between different chromatin molecules) of the condensed phase for each salt concentration. The resulting plot is in figure 6.5b, where we observe that the connectivity of the breathing chromatin is significantly higher than the connectivity of non-breathing chromatin (approximately double for equivalent salt concentrations). The increased connectivity means there are more inter-molecular attractive interactions, an increased number of interactions is essential to overcome the entropy loss of de-mixing. Thus we observe that the increased nucleosome valency, via nucleosome breathing, increases the liquid network connectivity which stabilizes (reduces the critical salt concentration) the LLPS of chromatin. The important role of the connectivity on the stability of more generic and diverse biomolecular condensates has been studied in [204].

6.3 Extrapolation to larger chromatin system sizes

We now use our minimal model to simulate larger system sizes: chromatin with 24, 50, 100, and 200 nucleosomes. We stick with the 165 NRL that we parameterized

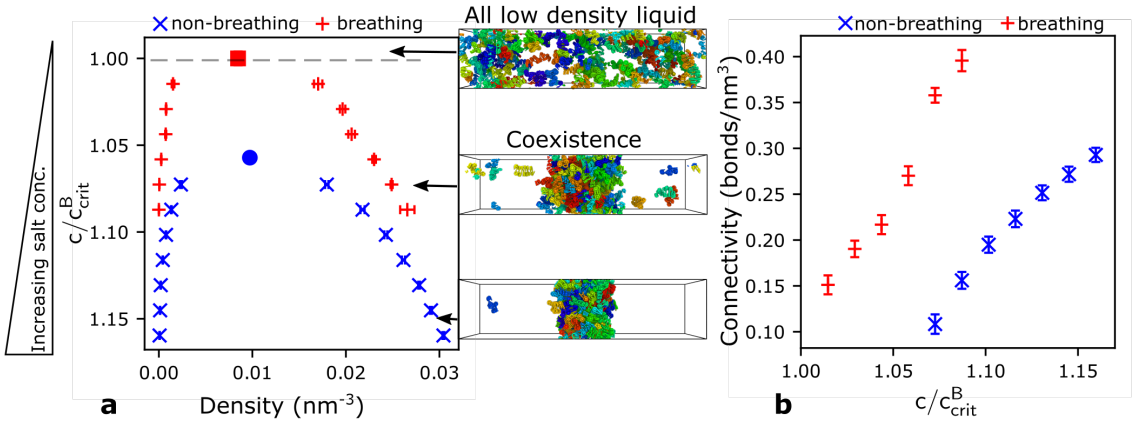

Figure 6.5: **Impact of nucleosome breathing on the phase behavior of chromatin. (a)** Liquid-liquid phase diagram of a solution of 12-nucleosome chromatin arrays in salt concentration (vertical axis) versus chromatin density space (horizontal axis, units of number density in nm^{-3}). The salt concentration c is normalized in terms of the critical salt concentration of the breathing chromatin c^B_{crit}, the axis is plotted such that increasing salt concentration goes downwards. The data points (blue: non-breathing, red: breathing) represent coexistence points, i.e. the densities of the dilute (left branch) and condensed (right branch) phases in coexistence with each other. The critical points (blue circle: non-breathing, red square: breathing) are calculated from the data points using equations 6.7 and 6.8. Above the critical point is the one-phase region where chromatin exists as a well mixed dilute liquid, where the DNA–DNA repulsion dominates due to the low salt concentration. Below the critical point is the two phase region where liquid-liquid phase separation occurs spontaneously, i.e if the salt concentration is increased above c_{crit} the chromatin will de-mix into a condensate and dilute liquid phase. The simulation snapshots indicate typical configurations at the indicated locations in the phase diagram. The absolute value of c^B_{crit} is 0.069 mol/L and c^{NB}_{crit} is 0.073 mol/L. **(b)** The connectivity of chromatin (inter-chromatin bonds per unit volume) in the condensate for non-breathing (blue points) and breathing (red) nucleosomes.

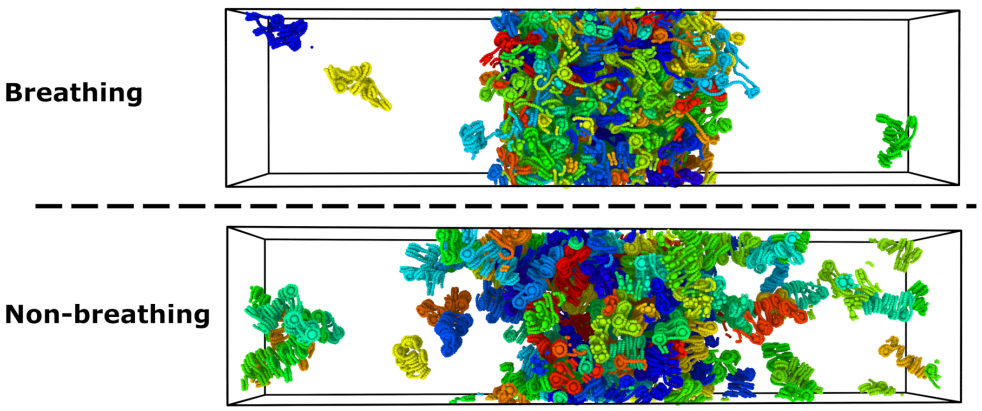

Figure 6.6: **Snapshots of breathing and non-breathing chromatin coexistence simulations at the same salt concentration.**

the model for and simulate at four salt concentrations: 0.05 0.07, 0.1, and 0.15 mol/L NaCl which we found was a suitable range to cover all levels of chromatin compaction in the chemically-specific model.

6.3.1 Methods

We generated the initial structures using the version 2 method in section 5.4. This means that the breathing and non-breathing chromatin have the same initial configuration but differ in the amount of DNA rigidly bound to the nucleosome core. At each salt concentration, for each system size, for breathing and non-breathing, we ran 20 repeats for approximately 50 million timesteps each. Simulation snapshots were recorded every 100,000 timesteps and the second half of the trajectories were used for used analysis, resulting in approximately 5000 frames for each data point. We computed the sedimentation coefficients using the following equation as done in previous coarse-grained chromatin work [16, 20]:

$$S = S_1 \left(1 + \frac{2R_1}{N} \sum_i \sum_{j<i} \frac{1}{R_{ij}} \right), \tag{6.9}$$

where S_1 is the sedimentation coefficient of a single nucleosome (set to 11.1 S), R_1 is the radius of a nucleosome (set to 55 Å), N is the number of nucleosomes, and R_{ij} are the inter-nucleosome distances.

We calculated the inter-nucleosome contact matrices as:

$$C_{ij} = \frac{1}{N_t} \sum_t C_{ij}(t), \tag{6.10}$$

where $C_{ij}(t)$ is the inter-nucleosome contact matrix at trajectory frame t and N_t is the total number of frames used in the analysis.

$$C_{ij}(t) = \begin{cases} 1, & \text{if distance between nucleosomes } i \text{ and } j < 110\text{Å} \ , \\ 0, & \text{otherwise.} \end{cases} \tag{6.11}$$

6.3.2 Results and discussion

The mean values of S are plotted in figure 6.7 along with the computed histograms drawn as split-violin plots at the corresponding data points. The general trend is the same for all system sizes, at the lowest salt, breathing chromatin is less condensed (lower sedimentation coefficient) than non-breathing chromatin, whilst at the highest salt breathing chromatin is more condensed (larger sedimentation coefficient) than non-breathing chromatin. The cross-over point of the curves changes as the system size increases, for N=24 at 0.07 mol/L the mean S values of breathing and non-breathing chromatin are very similar, whilst the breathing distribution extends to lower S values than the non-breathing distribution. As N is increased to 50, the mean value of S for breathing is slightly larger than for non-breathing at 0.07 mol/L. However, the distribution still overlaps with the non-breathing distribution on both sides. For N=200, the mean value of S for breathing is significantly larger, completely outside the non-breathing distribution range and the breathing distribution does not overlap the non-breathing distribution for lower S.

At the lowest salt value (0.05 mol/L) the overall chromatin structure does not change as N increases, they remain as extended fibers, breathing chromatin is more extended than the non-breathing and the fibers are self-similar—i.e. if we were to cut out a 24-N section from the 200-N chromatin it would look the same as the 24-N chromatin. For 0.1 and 0.15 mol/L at all N the breathing chromatin condenses into a liquid-like, approximately spherical, droplet. This liquid-like behavior directly follows from the chemically-specific model where we found that nucleosome breathing produces liquid-like chromatin.

The non-breathing chromatin exhibits more complex behavior as salt is increased and N changes. At 0.15 mol/L N=24 we observe the zig-zag ladder structures similar to as observed for the 12-nucleosome chemically-specific model. Looking at the sedimentation coefficient distributions for N=24, from 0.15 to 0.07 mol/L we see little difference, just a small decrease as the salt is lowered, very similar to the chemically-specific model curves in figure 4.11a. However, when we look at N=50 we observe that at 0.15 mol/L the sedimentation coefficient distribution is bi-modal: the lower peak ($S \approx 100$ S) corresponds to a standard 50-N fully-stacked but linear zig-zag fiber, while the upper peak ($S \approx 130$ S) corresponds to a zig-zag fiber that is folded in half onto itself. These two configurations are illustrated in figure 6.7b right side. At 0.1 mol/L the distribution is mostly centered on the linear-zig-zag ($S \approx 100$) but the bi-modality is still present with a smaller but significant population of folded over fibers. Even at 0.07 mol/L, the tail of the non-breathing distribution reaches 130 S indicating there is a non-zero occurrence of the folded zig-zags. At N=100 the distributions are long-tailed, the peaks at $S \approx 220$ S correspond to high density folded up structures, nearing the compaction levels of the breathing chromatin, while the long tails show that the fibers occasionally unfold to lower compaction levels. The distribution for 0.1 mol/L N=100 has a particularly long tail, the end of which ($S \approx 110$) corresponds to a fully linear-zig-zag, which is annotated in figure 6.7c. For N=200 the 0.15 mol/L distribution looks Gaussian, corresponding to chromatin always being in a compact folded up state. The 0.1 mol/L distribution has a medium length tail, the chromatin can exist in states that are not completely compacted but it never extends into the fully unfolded linear-zig-zag state, this can only be reached by decreasing the salt concentration. These data suggests that even when breathing motions are restricted, large-scale chromatin arrays (i.e., with 200 nucleosomes or more) are unlikely to condense into the long and ordered zig-zag configurations proposed in textbooks; indeed, such long fiber-like structures have been observed via cryo-EM of reconstituted chromatin with less than 80 nucleosomes [190]. This results put forward an additional parameter that destabilized the folding of chromatin into ordered fibers: increasing number of nucleosomes.

We next computed the inter-nucleosome contact maps for each system at 0.15 mol/L, which are plotted in figure 6.8. The contact map for the non-breathing chromatin is plotted on the upper triangle (in blue) and that for the breathing chromatin is plotted on the lower triangle (in red). A contact matrix is symmetric, so no information is lost by only plotting one triangle. We see that the non-breathing matrices have a very distinct pattern of a high-intensity 2nd diagonal — this is the k=2 face–face stacking of nucleosomes in a zig-zag arrangement. These values are the largest, reaching a frequency of 0.6 which means for 60% of the trajectories that specific contact exists. The maximum frequency in the non-breathing trajectories

is 0.3. We observe that the rest of the matrix is non-zero with a small interaction frequency, this is due to folding over of the zig-zag fibers while still maintaining most of the face–face k=2 interactions. The breathing chromatin matrices have the maximum on the k=3 diagonal, this is different to the 12-N chemically-specific breathing simulations which had the maximum at k=2. We also see that the breathing matrix has larger frequencies further away from the main diagonal, these are contacts occurring between nucleosomes separated by many neighbors facilitated by the liquid-like droplet configuration of breathing chromatin at 0.15 mol/L salt. Hence, variations in the linker DNA length, in this case via breathing motions, facilitate the formation of long-range chromatin interactions, which might be important to gene regulation. Because in vivo, chromatin parameters are highly heterogeneous (e.g. the linker DNA are not uniform, but vary from nucleosome to nucleosome, there are chemical modifications across different histones, and a wide-range of proteins can target chromatin), the structural behavior of chromatin in cells is more likely to resemble the liquid-like organization that we observe here for breathing nucleosomes, than the zig-zag fiber-like organization.

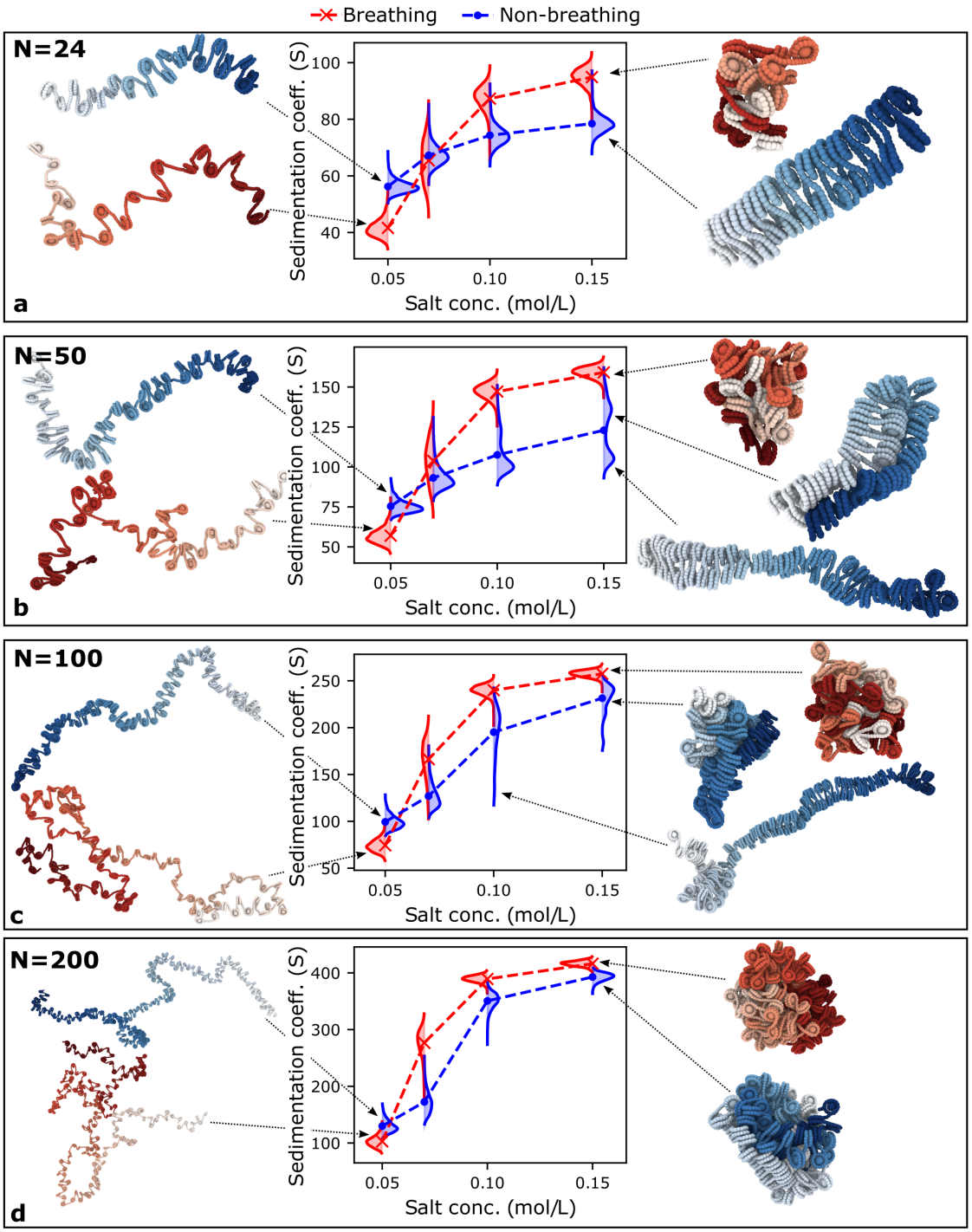

Figure 6.7: **N-nucleosome length chromatin. (a, b, c, d)** Sedimentation coefficients of chromatin with 24, 50, 100, and 200 nucleosomes respectively as a function of monovalent salt concentration. The blue lines are for non-breathing chromatin and the red lines are for breathing chromatin. The data points are the mean values, the full distributions (normalized histograms) of the sedimentation coefficients are plotted at the corresponding mean value as split-violin plots with breathing on the left in red and non-breathing on the right in blue. The simulation snapshots illustrate typical chromatin configurations at the labeled locations on the plots. The color-scheme of the snapshots is consistent with the plots: red is breathing, blue is non-breathing. The color gradient indicates the nucleosome numbering.

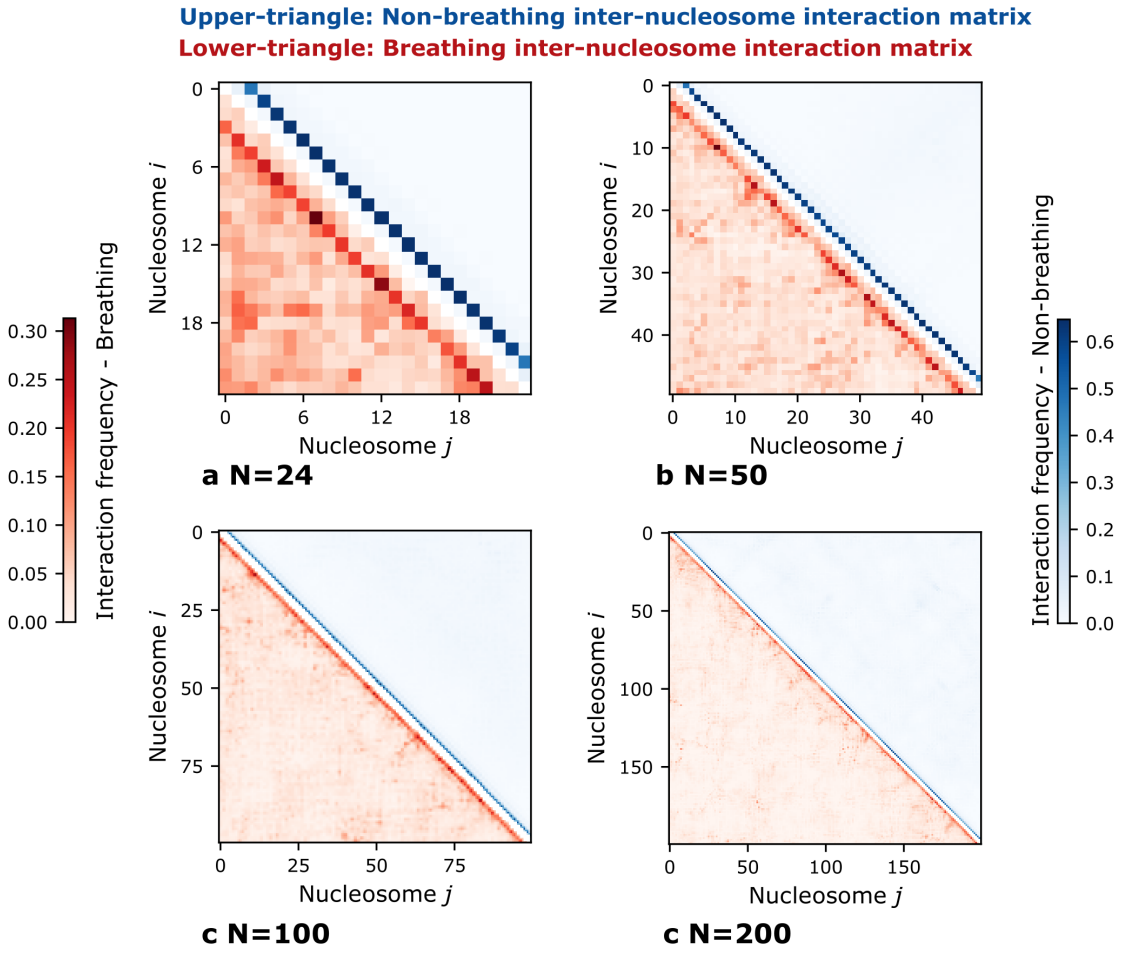

Figure 6.8: **Inter-nucleosome contact matrices for N-nucleosome length chromatin**. **(a, b, c, d)** Inter-nucleosome contact matrices at 0.15 mol/L for chromatin with 24, 50, 100, and 200 nucleosomes respectively. The upper triangle is for non-breathing chromatin (colored blue) and the lower triangle is for breathing chromatin (colored red). The color scales (interaction frequency) give the intensity of the contacts, the number is the proportion of time nucleosomes i and j are in contact, 0 means never in contact, 1 means in contact 100% of the time.

6.4 Periodicity in chromatin compaction for regular NRLs

We have used the minimal model to investigate the previously reported periodicity in chromatin compaction for different regular NRLs [212]. We computed the radius of gyration of each NRL in the range 160 bp to 211 bp for 12-nucleosome and 24-nucleosome non-breathing chromatin.

6.4.1 Methods

To use the minimal model for NRLs that are not a multiple of 5, we use the following method: First we use the chemically specific model to create an initial structure with the desired NRL, we then map the structure to the minimal model representation by grouping every sets of 5 DNA base-pairs. However, when we reach the very end of the chromatin, we neglect any last base pairs that remain. To be explicit for 12-nucleosome 167 NRL, we generate a chemically-specific structure that has $167 \times 12 = 2004$ bp. Dividing this by 5, we have 400 segments of length 5 bp and a remainder of 4 extra base pairs. Therefore, we map the first 2000 bp, counting from one end of the DNA, to 400 minimal model DNA beads, we then ignore the last 4 bp. The minimal DNA beads are then categorized into linker DNA or nucleosomal DNA as before.

We simulated at 0.15 mol/L and used temperature replica exchange molecular dynamics (T-REMD) with a temperature range of 300–600 K and 16 replicas for N=12 and 24 replicas for N=24. The simulations were run for 10 million timesteps, with snapshots recorded every 5000 timesteps. We removed the first 1 million timesteps of the trajectories from the analysis. This resulted in 1800 coordinate snapshots for each NRL. The radius of gyrations were computed using just the coordinates of the core beads. To get a representative snapshot for each simulation, we performed clustering analysis of the RMSD of particle coordinates using the method of Daura et al [213]. The snapshots in figure 6.9 and figure 6.10 are the snapshots with the most neighbors with a RMSD difference of less than 50 Å.

6.4.2 Results and discussion

The radius of gyration for each NRL is shown in figure 6.9, the green data points are for 12-nucleosome chromatin arrays and the purple points are for 24-nucleosome chromatin arrays. We see a clear periodicity in the radius of gyration, in reasonable agreement with the pattern seen by Zhurken et al [212] from their MC simulations and experimental data. That is, chromatin is more compact for the 147+10n NRLs and less compact for the 147+10n+5 NRLs where N is an integer. For example, the data point for 167 NRL 12N has a Rg of 115Å, while 172 has a Rg of 130 Å. The periodicity we observe is not exactly 10 bp, but rather ≈ 10.5, this is demonstrated by the peaks for the 12N curve: 163, 173, 184, 195, 207; and for the 24N curve: 164, 174, 186, 198, 209. The 10.5 bp periodicity arises from the average twist between DNA base-pairs which is 34 degrees. An interesting observation is that the 12N and 24N curves are not completely aligned, rather the peaks for 24N lag by 1 to 2 bp from the 12N peaks. The reason for this is the same length dependent behavior we saw in figure 6.7; that is, certain extended ladder-like structures that are stable

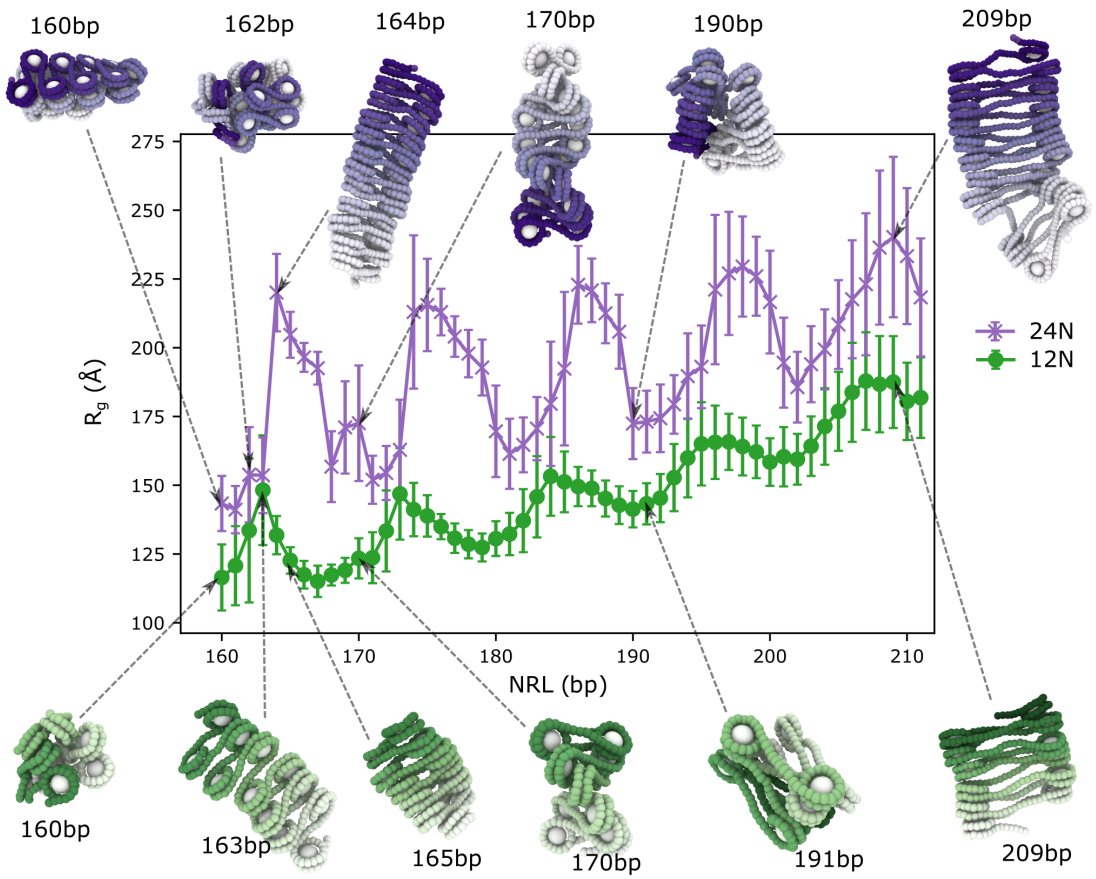

Figure 6.9: **Periodicity in chromatin compaction for regular NRLs.** Purple crosses are for 24-nucleosome chromatin, green circles are for 12-nucleosome chromatin. The data points and errorbars are the mean ± standard deviation of the radius of gyration of the core beads. Representative configurations for select data points are shown, the NRL is given in units of bp (base-pair), the configurations were chosen by clustering analysis.

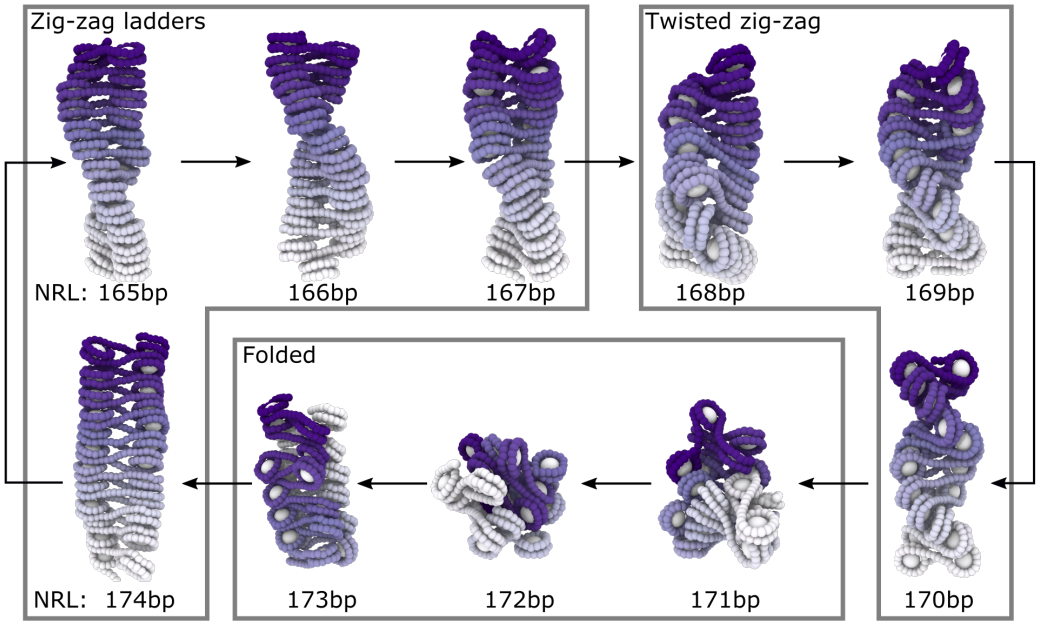

Figure 6.10: **Chromatin structures arising from different NRLs**. The structures can be categorized into 3 main categories: zig-zag ladders, twisted zig-zags, and folded which occur following a cyclic pattern within the ∼10 bp periodicity.

for 12N will fold-over when the system size increases to 24N. An example of this behavior occurs in the 163 NRL system. The 12N 163 NRL structure, pictured in figure 6.9, is a loosely stacked zig-zag because the torsional restraints imposed by the DNA forces successive nucleosomes to be rotated by a few degrees with respect to one another, hence hindering them from stacking perfectly flat on top of each other. When the number of nucleosomes in the array increase from 12 to 24 in this 163 NRL system, the slight nucleosome rotations that prevent flat stacking, imply that it is more energetically favorable for the fiber to fold over itself. In contrast, in the 165 NRL case, the length of the DNA instead completes almost a full turn, which allows successive nucleosomes to stack almost completely flat on top of each other. Consistently, the most extended 24N structure is the 164 NRL, which, as pictured, also has more regular/flat nucleosome stacking. The structures can be grouped into 3 main categories: Zig-zag ladders, twisted zig-zags, and folded over structures, which occur periodically as shown in figure 6.10.

6.5 Inter-chromatin fiber interactions

Experiments by Gibson et al [31] have found that the LLPS of chromatin in vitro is modulated by the chromatin nucleosome repeat length. Specifically, they found that NRLs of 147+10n+5 exhibit phase separation at lower salt concentrations than NRLs of 147+10n, or in other words the range of stability of LLPS is larger for NRLs of 147+10n+5 than 147+10n, where n is an integer. To investigate this phenomenon and gain an understating of the molecular level reasons behind it, we computed the inter-chromatin PMFs for 12-nucleosome chromatin fibers for five different NRLs. Three 147+10n+5 fibers (162,172,182), and two 147+10n fibers (167,177).

6.5.1 Methods

We computed the PMFs at 0.15 mol/L NaCl using umbrella sampling with the COLVARS package [162] where the collective variable the distance between the centers of mass of the two chromatin fibers (R). The centers of mass are computed using only the core beads. Additionally, we used T-REMD for each umbrella window to enhance the sampling. This was necessary due to the large number of different chromatin configurations that can occur with the same R. Importantly, the T-REMD was done independently for each window, i.e no exchanges occur between different windows, simply each window was run using a temperature replica exchange scheme with 16 replicas spanning from 300–600 K. The standard metropolis acceptance criteria can be used as long as we are careful to add the biasing potential to the total potential energy of the system. The collective variable R was biased using a harmonic potential with a force constant of 0.002 kcal/mol/Å^2 and 16 equally spaced umbrella windows were used spanning 0–600 Å. Each window was run for 10 million timesteps. The PMFs were computed from the umbrella windows using WHAM [97]. As R is a 3D distance the output from WHAM is the free energy curve $F = -k_BT\log(P(R))$ where P is the probability. To convert to the PMF curve $\text{PMF} = -k_BT\log(p(R))$ where p is the probability density we include the Jacobian term which is the volume accessible at a distance of R:

$$J = 4\pi R^2 \Delta R, \tag{6.12}$$

$$\text{PMF}(R) = F(R) + k_BT\log(J(R)), \tag{6.13}$$

where ΔR is the grid spacing used in the WHAM calculation. This was not needed for any of our previous umbrella sampling calculations as they were 1D collective variables where $P \propto p$.

6.5.2 Results and discussion

The PMF curves are plotted in figure 6.11a. We observe than both 167 and 177 NRL have shallower minima than 162, 172, and 182 NRL. This means that the inter-fiber chromatin interactions, at physiological salt, are weaker between 147+10n NRL chromatin than they are for 147+10n+5 NRL chromatin. This agrees with the results of Gibson et al [31] that 147+10n+5 NRL have larger range of LLPS stability than 147+10n, which suggests stronger inter-fiber chromatin interactions for 147+10n+5 than 147+10n NRLs. Representative simulation snapshots corresponding to the free energy minima for each system are shown in figure 6.11b and c. We observe that the 10n systems have few inter-chromatin contacts and remain in their respective single molecule equilibrium states — the regular zig-zag ladders, as pictured in figure 6.10. This is in contrast to the 10n+5 systems which have multiple inter-chromatin contacts facilitated by the disordered single fiber configurations, e.g. 172 NRL in figure 6.10. We postulate that this balance between inter and intra-chromatin interactions is the key reason for the observed differences in the PMFs and the LLPS observed in experiment [31]. The regular zig-zag fiber structures of the 147+10n chromatin systems are characterized by stronger intra-chromatin face-to-face nucleosome stacking interactions than the 147+10n+5 counterparts. Thus, the free-energy of inter-fiber chromatin interactions we measure in the PMFs of the 147+10n systems are mostly contributed by weaker side–side

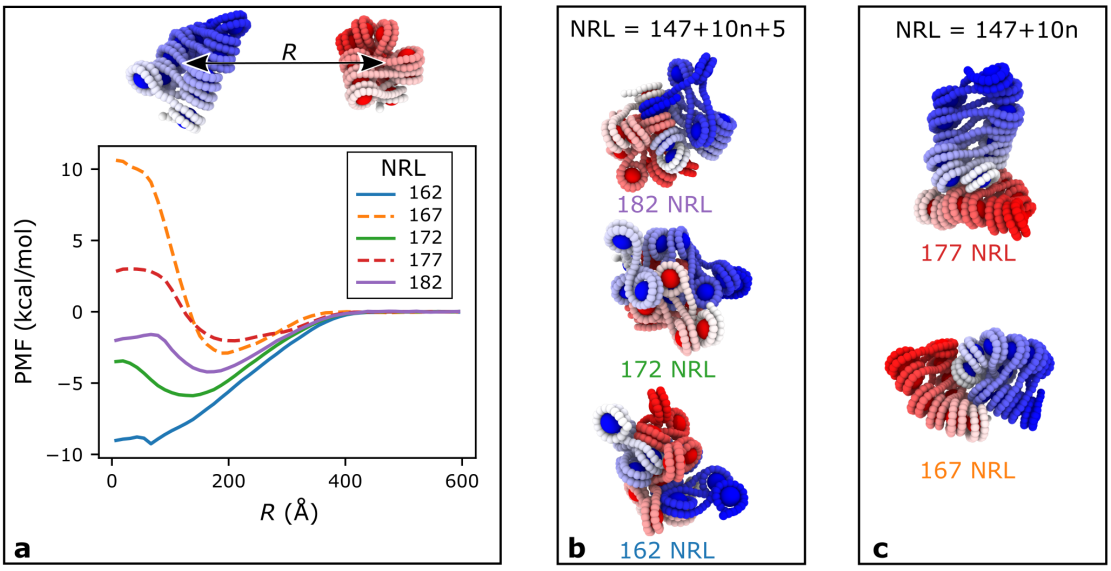

Figure 6.11: **PMFs between two 12-nucleosome chromatin fibers for different NRLs**. **(a)** PMFs for 162, 167, 172, 177, and 182 NRL. The 147+10n+5 NRLs are solid lines, the 147+10n are dashed lines. The upper panel illustrates the collective variable R — the distance between the centers of mass of the two chromatin fibers. **(b)** Snapshots of representative chromatin configurations that occur at the PMF minima for each 147+10n+5 NRL. **(c)** Snapshots of representative chromatin configurations that occur at the PMF minima for each 147+10n NRL.

nucleosome–nucleosome interactions. In contrast, the 147+10n+5 chromatin arrays are more irregularly folded, which enables them to form strong nucleosome face-to-face stacking interactions between the two fibers, thus the overall inter-fiber chromatin interaction measured by the PMFs are stronger.

Chapter 7

Conclusions and outlook

In this work, we have developed a multiscale model of chromatin that spans the resolutions and system sizes appropriate to link the fine atomistic details of nucleosomes to the emergence of chromatin self-organization and LLPS in systems with over a thousand nucleosomes.

Our chemically-specific model is at residue/base-pair resolution, with this we have been able to investigate:

- Nucleosome formation. We found that DNA supercoiling and torsional restraints are required for reliable formation of nucleosomes, while the chirally inverted reversome can be created by inverting the supercoiling.

- The unwrapping behavior of single nucleosomes. We computed the free energy curves of the DNA extensions and the rupture forces for the transitions between the 3 main nucleosome unwrapping states. Our values were in good agreement with experiments. Furthermore, we found that DNA sequence can have a noticeable effect on the free energy of nucleosome unwrapping, that high-salt encourages nucleosome unwrapping and low salt hinders it, that removing histones tails significantly destabilizes the fully wrapped state of nucleosomes with the partially unwrapped state becoming the free energy minimum, and that the binding of the H1 linker histone protein increases the force needed to transition from the fully wrapped state into the partially wrapped state.

- Force extension of chromatin. We computed the force extension curves of 4-nucleosome chromatin arrays using steered molecular dynamics, observing the typical saw-tooth pattern characteristic of each nucleosome suddenly unwrapping during the pulling experiments. The measured forces were in agreement with experimental values. We further found that the addition of H1 linker histone significantly increases the forces needed to unwrap the first nucleosome turn, and also increases the forces needed to reach the nucleosome rupture regime.

- Nucleosome sliding. We computed the diffusion coefficient of nucleosome sliding for the high affinity NCP147 sequence and low affinity poly-A sequence, finding a factor of 26 difference. Furthermore, we observed two different sliding mechanisms consistent with the continuous twist diffusion and the discrete loop propagation motions, as observed in previous work.

- The structure of 12-nucleosome chromatin as a function of the monovalent salt concentration for breathing and non-breathing chromatin. We found that the compaction of chromatin increases as the monovalent salt concentration increases and our values of the sedimentation coefficients at 0.15 mol/L were in agreement with experiment values. For the short NRL of 165 bp, we found that non-breathing nucleosomes result in a zig-zag fiber chromatin configuration, consistent with structures observed in vitro and in previous computational work. We found that allowing the nucleosomes to breath (spontaneous DNA unwrapping/sliding due to thermal motions) resulted in destabilization of the zig-zag fiber creating denser structures at high salt, and more open structures at low salt; this is in agreement with the liquid-like nature of chromatin proposed by Maeshima and colleagues. Our work proposes that the liquid-like nature can arise from the plasticity of the nucleosomes; that is, nucleosomes in solution and in cellss are not completely rigid structures, but exhibit dynamic structural fluctuations. We additionally found that including H1 linker histone destabilizes the zig-zag ladder chromatin structure resulting in significantly more compact structures.

Our minimal model represents nucleosomes with a single bead for the core protein and a bead for every five DNA base-pairs, this enables larger scale chromatin simulations while still incorporating the important features of the bending/torsional rigidity, the excluded volume size of the DNA, and the orientation dependent nucleosome–nucleosome interactions. With our minimal model we have investigated:

- Liquid-liquid phase separation of chromatin. We computed the LLPS phase diagrams of 12-nucleosome 165 NRL chromatin as a function of monovalent salt for breathing and non-breathing nucleosomes. We found that the critical salt concentration is lower when the nucleosomes are allowed to breath, thus the range of stability of the LLPS coexistence region is increased for breathing nucleosomes. This proposes that the plasticity of the nucleosomes, which gives rise to the liquid-like nature of chromatin, can enhance the LLPS of chromatin, and hence, contributes to the regulation of the organization and membraneless compartmentalization of the genome.

- Larger chromatin system sizes. We investigated the scaling behavior of isolated chromatin fibers as the number of nucleosomes is increased from 24 to 200. We found the overall salt dependent behavior is consistent for all systems, regardless of the number of nucleosomes; that is, as the salt concentration is increased the chromatin compaction increases. Furthermore, the structure of breathing chromatin exhibits a greater range of compaction: it is more open at low salt and more condensed at high salt compared to non-breathing nucleosomes. At high salt the breathing chromatin condenses into liquid-like droplets similarly for all system sizes, whereas the non-breathing chromatin exhibits nucleosome-number dependent behavior—when the number of nucleosomes is less than 50 the zig-zag ladder structures remains stable, when it is greater than 50 the zig-zag ladder folds over and the fully extended ladder configurations are no longer energetically favorable. This behavior is controlled by the balance between a few strong face–face nucleosome–nucleosome stacking interactions and multiple weaker side–side nucleosome–nucleosome interactions.

- Periodicity in chromatin compaction for different NRLs. We computed the radius of gyration and equilibrium structures for 12-nucleosome and 24-nucleosome chromatin for all NRLs in the range 160–211 bp. We found the compaction levels and structural types follow a periodic pattern with a periodicity of approximately 10.5 bp, which arises from the mean twist between DNA base-pairs of approximately 34 degrees. We found a slight difference of 1-4bp for 24N vs 12N chromatin — the most extended ladder structures for 12N are not stable for 24N and fold over. The chromatin structures we observe can be classed into 3 main categories: Zig-zag ladder, twisted zig-zag, and folded.

- Inter-chromatin interactions. We computed the PMFs for the interaction between two 12-nucleosome chromatin fibers. We found that chromatin fibers with NRLs of 147+10n+5 have significantly stronger inter-chromatin interactions than NRLs of 147+10n, where n is an integer. This result is in agreement with experiments that found 147+10n+5 NRL chromatin can exhibit LLPS at lower salt concentrations than 147+10n. The reason behind this behavior is the balance between the inter and intra-chromatin interactions. The 147+10n chromatin has more regular zig-zag fiber configurations with stronger intra-chromatin interactions and weaker inter-chromatin interactions. We further found a trend that as the NRL is increased the free energy minima become less deep, this is due the the extra DNA–DNA repulsion for longer NRLs.

In addition to the results we have presented, the models and methods we have developed in this work are a significant contribution to the chromatin computational modeling community. Our rigid-base-pair model with additional phosphate charges is the first of its kind to be implemented in a established MD code engine such as LAMMPS. In general, the rigid-base-pair model is used in Monte-Carlo simulations, by extending it to LAMMPS, we have facilitated its transferability to other different systems and its usage by the wider community, as well as benefiting from the parallel scalability and flexibility of the MD method. Our rigid-base-pair DNA model is transferable to other DNA–protein systems, this is demonstrated in section B.1, and the corresponding publication [214] where I applied the model to study a DNA binding blood protein. The more generalized minimal-RBP model is an important new model, implemented in LAMMPS, for simulating twistable-semi-flexible polymers for which where are only a few currently implemented solutions [215]. Similarly, the anisotropic potential provides a more customizable form of the Gay-Berne potential which could find uses for other coarse-grained applications.

Future work using the models developed in this work include further categorizing the zig-zag fiber configurations as a function of nucleosome number, NRL, and DNA flexibility. Computing LLPS phase diagrams for different NRLs and chromatin with more nucleosomes, e.g. we expect 24-nucleosome and 48-nucleosome systems should be possible. Following on from the discussion about our qualitative vs quantitative salt-dependent behavior of the chemically-specific model, we aim to more fully parameterize the electrostatic interactions, possibly using reduced charges according to counterion condensation theory [216, 217], as done in other work [70, 218].

Bibliography

[1] S. I. S. Grewal and S. Jia, Nature Reviews Genetics **8**, 35 (2007).

[2] B. Alberts, D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential Cell Biology, Fourth Edition* (Taylor & Francis Group, 2013).

[3] G. Ozer, A. Luque, and T. Schlick, Current Opinion in Structural Biology **31**, 124 (2015).

[4] C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder, and T. J. Richmond, Journal of Molecular Biology **319**, 1097 (2002).

[5] T. J. Richmond and C. A. Davey, Nature **423**, 145 (2003).

[6] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond, Nature **389**, 251 (1997).

[7] J. L. Compton, M. Bellard, and P. Chambon, Proc Natl Acad Sci U S A **73**, 4382 (1976).

[8] C. L. Woodcock, A. I. Skoultchi, and Y. Fan, Chromosome Research **14**, 17 (2006).

[9] D. J. Tremethick, Cell **128**, 651 (2007).

[10] K. Maeshima, R. Imai, S. Tamura, and T. Nozaki, Chromosoma **123**, 225 (2014).

[11] K. Maeshima, S. Tamura, J. C. Hansen, and Y. Itoh, Current Opinion in Cell Biology **64**, 77 (2020).

[12] J. C. Hansen, Annual Review of Biophysics and Biomolecular Structure **31**, 361 (2002).

[13] M. Kruithof, F. T. Chien, A. Routh, C. Logie, D. Rhodes, and J. Van Noort, Nature Structural and Molecular Biology **16**, 534 (2009).

[14] T. Schalch, S. Duda, D. F. Sargent, and T. J. Richmond, Nature **436**, 138 (2005).

[15] F. Song, P. Chen, D. Sun, M. Wang, L. Dong, D. Liang, R. M. Xu, P. Zhu, and G. Li, Science **344**, 376 (2014).

[16] O. Perišić, R. Collepardo-Guevara, and T. Schlick, Journal of Molecular Biology **403**, 777 (2010).

[17] S. A. Grigoryev, G. Arya, S. Correll, C. L. Woodcock, and T. Schlick, Proceedings of the National Academy of Sciences of the United States of America **106**, 13317 (2009).

[18] K. Maeshima, S. Ide, K. Hibino, and M. Sasai, Current Opinion in Genetics and Development **37**, 36 (2016).

[19] K. Maeshima, S. Hihara, and M. Eltsov, Current Opinion in Cell Biology **22**, 291 (2010).

[20] R. Collepardo-Guevara and T. Schlick, Proceedings of the National Academy of Sciences of the United States of America **111**, 8061 (2014).

[21] S. A. Grigoryev, G. Bascom, J. M. Buckwalter, M. B. Schubert, C. L. Woodcock, and T. Schlick, Proceedings of the National Academy of Sciences of the United States of America **113**, 1238 (2016).

[22] G. Bascom and T. Schlick, Biophysical Journal **112**, 434 (2017).

[23] M. A. Ricci, C. Manzo, M. F. García-Parajo, M. Lakadamyali, and M. P. Cosma, Cell **160**, 1145 (2015).

[24] S. A. Grigoryev and M. Schubert, Essays in Biochemistry **63**, 109 (2019).

[25] N. Krietenstein and O. J. Rando, Current Opinion in Genetics and Development **61**, 32 (2020).

[26] G. D. Bascom, T. Kim, and T. Schlick, Journal of Physical Chemistry B **121**, 3882 (2017).

[27] O. Wiese, D. Marenduzzo, and C. A. Brackley, Proceedings of the National Academy of Sciences of the United States of America **116**, 17307 (2019).

[28] A. R. Strom, A. V. Emelyanov, M. Mir, D. V. Fyodorov, X. Darzacq, and G. H. Karpen, Nature **547**, 241 (2017).

[29] A. G. Larson, D. Elnatan, M. M. Keenen, M. J. Trnka, J. B. Johnston, A. L. Burlingame, D. A. Agard, S. Redding, and G. J. Narlikar, Nature **547**, 236 (2017).

[30] F. Erdel and K. Rippe, Biophysical Journal **114**, 2262 (2018).

[31] B. A. Gibson, L. K. Doolittle, M. W. G. Schneider, L. E. Jensen, N. Gamarra, L. Henry, D. W. Gerlich, S. Redding, and M. K. Rosen, Cell **179**, 470 (2019).

[32] B. R. Sabari, A. Dall'Agnese, A. Boija, I. A. Klein, E. L. Coffey, K. Shrinivas, B. J. Abraham, N. M. Hannett, A. V. Zamudio, J. C. Manteiga, C. H. Li, Y. E. Guo, D. S. Day, J. Schuijers, E. Vasile, S. Malik, D. Hnisz, T. I. Lee, I. I. Cisse, R. G. Roeder, P. A. Sharp, A. K. Chakraborty, and R. A. Young, Science **361** (2018).

[33] D. Hnisz, K. Shrinivas, R. A. Young, A. K. Chakraborty, and P. A. Sharp, Cell **169**, 13 (2017).

[34] A. Boija, I. A. Klein, B. R. Sabari, A. Dall'Agnese, E. L. Coffey, A. V. Zamudio, C. H. Li, K. Shrinivas, J. C. Manteiga, N. M. Hannett, A. Dall'Agnese, E. L. Coffey, A. V. Zamudio, C. H. Li, K. Shrinivas, J. C. Manteiga, N. M. Hannett, B. J. Abraham, L. K. Afeyan, Y. E. Guo, J. K. Rimel, C. B. Fant, J. Schuijers, T. I. Lee, D. J. Taatjes, R. A. Young, A. Dall'Agnese, E. L. Coffey, A. V. Zamudio, C. H. Li, K. Shrinivas, J. C. Manteiga, and N. M. Hannett, Cell **175**, 1842 (2018).

[35] A. J. Plys, C. P. Davis, J. Kim, G. Rizki, M. M. Keenen, S. K. Marr, and R. E. Kingston, Genes and Development **33**, 799 (2019).

[36] Y. Zhang, B. Bertulat, A. H. Tencer, X. Ren, G. M. Wright, J. Black, M. C. Cardoso, and T. G. Kutateladze, iScience **17**, 182 (2019).

[37] W. K. Cho, J. H. Spille, M. Hecht, C. Lee, C. Li, V. Grube, and I. I. Cisse, Science **361**, 412 (2018).

[38] S. J. Nair, L. Yang, D. Meluzzi, S. Oh, F. Yang, M. J. Friedman, S. Wang, T. Suter, I. Alshareedah, A. Gamliel, Q. Ma, J. Zhang, Y. Hu, Y. Tan, K. A. Ohgi, R. S. Jayani, P. R. Banerjee, A. K. Aggarwal, and M. G. Rosenfeld, Nature Structural and Molecular Biology **26**, 193 (2019).

[39] M. Boehning, C. Dugast-Darzacq, M. Rankovic, A. S. Hansen, T. Yu, H. Marie-Nelly, D. T. McSwiggen, G. Kokic, G. M. Dailey, P. Cramer, X. Darzacq, and M. Zweckstetter, Nature Structural and Molecular Biology **25**, 833 (2018).

[40] S. Sanulli, M. J. Trnka, V. Dharmarajan, R. W. Tibble, B. D. Pascal, A. L. Burlingame, P. R. Griffin, J. D. Gross, and G. J. Narlikar, Nature **575**, 390 (2019).

[41] E. M. Hildebrand and J. Dekker, Trends in Biochemical Sciences **45**, 385 (2020).

[42] R. Hancock, International Review of Cell and Molecular Biology **307**, 15 (2014).

[43] R. Hancock, Frontiers in Physics **2** (2014), 10.3389/fphy.2014.00053.

[44] A. R. Strom and C. P. Brangwynne, Journal of Cell Science **132** (2019), 10.1242/jcs.235093.

[45] A. A. Hyman, C. A. Weber, and F. Jülicher, Annual Review of Cell and Developmental Biology **30**, 39 (2014).

[46] S. Alberti, A. Gladfelter, and T. Mittag, Cell **176**, 419 (2019).

[47] J. Dekker, M. A. Marti-Renom, and L. A. Mirny, Nature Reviews Genetics **14**, 390 (2013).

[48] S. Rao, M. Huntley, N. Durand, E. Stamenova, I. Bochkov, J. Robinson, A. Sanborn, I. Machol, A. Omer, E. Lander, and E. Aiden, Cell **159**, 1665 (2014).

[49] T.-H. Hsieh, A. Weiner, B. Lajoie, J. Dekker, N. Friedman, and O. Rando, Cell **162**, 108 (2015).

[50] H. D. Ou, S. Phan, T. J. Deerinck, A. Thor, M. H. Ellisman, and C. C. O'Shea, Science **357**, eaag0025 (2017).

[51] J. Widom, Quarterly Reviews of Biophysics **34**, 269–324 (2001).

[52] M. Y. Tolstorukov, A. V. Colasanti, D. M. McCandlish, W. K. Olson, and V. B. Zhurkin, Journal of Molecular Biology **371**, 725 (2007).

[53] S. Geggier and A. Vologodskii, Proceedings of the National Academy of Sciences of the United States of America **107**, 15421 (2010).

[54] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J.-P. Z. Wang, and J. Widom, Nature **442**, 772 (2006).

[55] N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, E. M. LeProust, T. R. Hughes, J. D. Lieb, J. Widom, and E. Segal, Nature **458**, 362 (2008).

[56] S. C. Satchwell, H. R. Drew, and A. A. Travers, Journal of Molecular Biology **191**, 659 (1986).

[57] R. Wu and H. Li, Genome Research **20**, 473 (2010).

[58] K. Struhl and E. Segal, Nature Structural &Amp; Molecular Biology **20**, 267 (2013).

[59] E. Segal and J. Widom, Current Opinion in Structural Biology **19**, 65 (2009), folding and binding / Protein-nuclei acid interactions.

[60] W. Iwasaki, Y. Miya, N. Horikoshi, A. Osakabe, H. Taguchi, H. Tachiwana, T. Shibata, W. Kagawa, and H. Kurumizaka, FEBS Open Bio **3**, 363 (2013).

[61] N. Happel and D. Doenecke, Gene **431**, 1 (2009).

[62] C. A. R. and H. J. J., FEBS Letters **589**, 2914 (2015).

[63] J. Bednar, I. Garcia-Saez, R. Boopathi, A. R. Cutter, G. Papai, A. Reymer, S. H. Syed, I. N. Lone, O. Tonchev, C. Crucifix, H. Menoni, C. Papin, D. A. Skoufias, H. Kurumizaka, R. Lavery, A. Hamiche, J. J. Hayes, P. Schultz, D. Angelov, C. Petosa, and S. Dimitrov, Molecular Cell **66**, 384 (2017).

[64] A. Sridhar, S. E. Farr, G. Portella, T. Schlick, M. Orozco, and R. Collepardo-Guevara, Proceedings of the National Academy of Sciences **117**, 7216 (2020).

[65] P. D. Dans, J. Walther, H. Gómez, and M. Orozco, Current Opinion in Structural Biology **37**, 29 (2016), theory and simulation • Macromolcular machines.

[66] A. L. B. Pyne, A. Noy, K. H. S. Main, V. Velasco-Berrelleza, M. M. Piperakis, L. A. Mitchenall, F. M. Cugliandolo, J. G. Beton, C. E. M. Stevenson, B. W. Hoogenboom, A. D. Bates, A. Maxwell, and S. A. Harris, Nature Communications **12**, 1053 (2021).

[67] S. O. Yesylevskyy, L. V. Schäfer, D. Sengupta, and S. J. Marrink, PLOS Computational Biology **6**, 1 (2010).

[68] L. Darré, M. R. Machado, A. F. Brandner, H. C. González, S. Ferreira, and S. Pantano, Journal of Chemical Theory and Computation **11**, 723 (2015), publisher: American Chemical Society.

[69] P. Šulc, F. Romano, T. E. Ouldridge, L. Rovigatti, J. P. K. Doye, and A. A. Louis, The Journal of Chemical Physics **137**, 135101 (2012).

[70] D. M. Hinckley, G. S. Freeman, J. K. Whitmer, and J. J. de Pablo, The Journal of Chemical Physics **139**, 144903 (2013).

[71] J. J. Uusitalo, H. I. Ingólfsson, P. Akhshi, D. P. Tieleman, and S. J. Marrink, Journal of Chemical Theory and Computation **11**, 3932 (2015).

[72] W. K. Olson, A. A. Gorin, X. J. Lu, L. M. Hock, and V. B. Zhurkin, Proceedings of the National Academy of Sciences of the United States of America **95**, 11163 (1998).

[73] F. Lankaš, J. Šponer, J. Langowski, and T. E. Cheatham, Biophysical Journal **85**, 2872 (2003).

[74] A. Pérez, A. Noy, F. Lankas, F. J. Luque, and M. Orozco, Nucleic Acids Research **32**, 6144 (2004).

[75] P. D. Dans, A. Pérez, I. Faustino, R. Lavery, and M. Orozco, Nucleic Acids Research **40**, 10668 (2012).

[76] A. Hospital, I. Faustino, R. Collepardo-Guevara, C. González, J. L. Gelpí, and M. Orozco, Nucleic acids research **41**, W47 (2013).

[77] G. Arya and T. Schlick, The Journal of Physical Chemistry A **113**, 4045 (2009).

[78] H. Jian, A. V. Vologodskii, and T. Schlick, Journal of Computational Physics **136**, 168 (1997).

[79] G. D. Bascom and T. Schlick, Biophysical Journal **114**, 2376 (2018).

[80] R. Collepardo-Guevara, G. Portella, M. Vendruscolo, D. Frenkel, T. Schlick, and M. Orozco, Journal of the American Chemical Society **137**, 10205 (2015).

[81] Y. Fan, N. Korolev, A. P. Lyubartsev, and L. Nordenskiöld, PLOS ONE **8**, 1 (2013).

[82] N. Clauvelin, P. Lo, O. I. Kulaeva, E. V. Nizovtseva, J. Diaz-Montes, J. Zola, M. Parashar, V. M. Studitsky, and W. K. Olson, Journal of Physics: Condensed Matter **27**, 064112 (2015).

[83] B. Eslami-Mossallam, R. D. Schram, M. Tompitak, J. van Noort, and H. Schiessel, PLOS ONE **11**, 1 (2016).

[84] A. Fathizadeh, A. Berdy Besya, M. Reza Ejtehadi, and H. Schiessel, The European Physical Journal E **36**, 21 (2013).

[85] J. Lequieu, A. Córdoba, J. Moller, and J. J. De Pablo, Journal of Chemical Physics **150** (2019), 10.1063/1.5092976.

[86] W. Li, P. G. Wolynes, and S. Takada, Proceedings of the National Academy of Sciences **108**, 3504 (2011).

[87] G. S. Freeman, J. P. Lequieu, D. M. Hinckley, J. K. Whitmer, and J. J. de Pablo, Phys. Rev. Lett. **113**, 168101 (2014).

[88] J. Lequieu, A. Córdoba, D. C. Schwartz, and J. J. De Pablo, ACS Central Science **2**, 660 (2016).

[89] B. E. de Jong, T. B. Brouwer, A. Kaczmarczyk, B. Visscher, and J. van Noort, Biophysical Journal **115**, 1848 (2018).

[90] T. Brouwer, C. Pham, A. Kaczmarczyk, W.-J. de Voogd, M. Botto, P. Vizjak, F. Mueller-Planitz, and J. van Noort, Nucleic Acids Research (2021), 10.1093/nar/gkab058.

[91] J.-B. Boulé, J. Mozziconacci, and C. Lavelle, Journal of Physics: Condensed Matter **27**, 033101 (2014).

[92] A. Bendandi, S. Dante, S. R. Zia, A. Diaspro, and W. Rocchia, Frontiers in Molecular Biosciences **7**, 15 (2020).

[93] J. Moller and J. J. de Pablo, Biophysical Journal **118**, 2057 (2020).

[94] S. Plimpton, Journal of Computational Physics **117**, 1 (1995).

[95] X. Lu and W. K. Olson, Nucleic Acids Research **31**, 5108 (2003).

[96] A. Stukowski, Modelling and Simulation in Materials Science and Engineering **18**, 015012 (2009).

[97] A. Grossfield, "Wham: the weighted histogram analysis method, version 2.0.9," (2020).

[98] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, Computational science (Elsevier Science, 2001).

[99] N. Grønbech-Jensen and O. Farago, Molecular Physics **111**, 983 (2013).

[100] D. A. Case, T. E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K. M. Merz Jr., A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, Journal of Computational Chemistry **26**, 1668 (2005).

[101] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, Journal of Computational Chemistry **4**, 187 (1983).

[102] L. D. Schuler, X. Daura, and W. F. van Gunsteren, Journal of Computational Chemistry **22**, 1205 (2001).

[103] W. L. Jorgensen and J. Tirado-Rives, Journal of the American Chemical Society **110**, 1657 (1988), pMID: 27557051.

[104] J. Huertas and V. Cojocaru, Journal of Molecular Biology , 166744 (2020).

[105] J. Jung, W. Nishima, M. Daniels, G. Bascom, C. Kobayashi, A. Adedoyin, M. Wall, A. Lappala, D. Phillips, W. Fischer, C.-S. Tung, T. Schlick, Y. Sugita, and K. Y. Sanbonmatsu, Journal of Computational Chemistry **40**, 1919 (2019).

[106] M. G. Saunders and G. A. Voth, Annual Review of Biophysics **42**, 73 (2013), pMID: 23451897.

[107] D. Reith, M. Pütz, and F. Müller-Plathe, Journal of Computational Chemistry **24**, 1624 (2003).

[108] S. Tanaka and H. A. Scheraga, Macromolecules **9**, 945 (1976).

[109] W. G. Noid, The Journal of Chemical Physics **139**, 090901 (2013).

[110] R. C. Bernardi, M. C. Melo, and K. Schulten, Biochimica et Biophysica Acta (BBA) - General Subjects **1850**, 872 (2015), recent developments of molecular dynamics.

[111] Y. Sugita and Y. Okamoto, Chemical Physics Letters **314**, 141 (1999).

[112] J. Kästner, WIREs Computational Molecular Science **1**, 932 (2011).

[113] A. Laio and M. Parrinello, Proceedings of the National Academy of Sciences **99**, 12562 (2002).

[114] C. Tsallis and D. A. Stariolo, Physica A: Statistical Mechanics and its Applications **233**, 395 (1996).

[115] D. J. Earl and M. W. Deem, Phys. Chem. Chem. Phys. **7**, 3910 (2005).

[116] A. Patriksson and D. van der Spoel, Phys. Chem. Chem. Phys. **10**, 2073 (2008).

[117] G. Bussi, Molecular Physics **112**, 379 (2014).

[118] P. Liu, B. Kim, R. A. Friesner, and B. J. Berne, Proceedings of the National Academy of Sciences **102**, 13749 (2005).

[119] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, Journal of Computational Chemistry **13**, 1011 (1992).

[120] B. Roux, Computer physics communications **91**, 275 (1995).

[121] S. E. Farr, E. J. Woods, J. A. Joseph, A. Garaizar, and R. Collepardo-Guevara, Nature Communications **12**, 2883 (2021).

[122] A. N. Boettiger, B. Bintu, J. R. Moffitt, S. Wang, B. J. Beliveau, G. Fudenberg, M. Imakaev, L. A. Mirny, C.-t. Wu, and X. Zhuang, Nature **529**, 418 (2016).

[123] G. S. Manning, Biophysical Journal **91**, 3607 (2006).

[124] S. Diekmann, Journal of Molecular Biology **205**, 787 (1989).

[125] W. K. Olson, M. Bansal, S. K. Burley, R. E. Dickerson, M. Gerstein, S. C. Harvey, U. Heinemann, X.-J. Lu, S. Neidle, Z. Shakked, H. Sklenar, M. Suzuki, C.-S. Tung, E. Westhof, C. Wolberger, and H. M. Berman, Journal of Molecular Biology **313**, 229 (2001).

[126] B. D. Coleman, W. K. Olson, and D. Swigon, The Journal of Chemical Physics **118**, 7127 (2003).

[127] F. Lankas, O. Gonzalez, L. M. Heffler, G. Stoll, M. Moakher, and J. H. Maddocks, Phys. Chem. Chem. Phys. **11**, 10565 (2009).

[128] W. K. Olson, A. A. Gorin, X.-J. Lu, L. M. Hock, and V. B. Zhurkin, Proceedings of the National Academy of Sciences **95**, 11163 (1998).

[129] F. Lankaš, J. Šponer, P. Hobza, and J. Langowski, Journal of Molecular Biology **299**, 695 (2000).

[130] X.-J. Lu, M. E. Hassan, and C. Hunter, Journal of Molecular Biology **273**, 668 (1997).

[131] X. Lu and W. K. Olson, Nucleic Acids Research **31**, 5108 (2003).

[132] X.-J. Lu and W. K. Olson, Nature Protocols **3**, 1213 (2008).

[133] G. L. Dignon, W. W. Zheng, Y. C. Kim, R. B. Best, and J. Mittal, Plos Computational Biology **14**, e1005941 (2018).

[134] I. Bahar, A. R. Atilgan, and B. Erman, Folding and Design **2**, 173 (1997).

[135] Y. C. Kim and G. Hummer, Journal of Molecular Biology **375**, 1416 (2008).

[136] S. Miyazawa and R. L. Jernigan, Journal of Molecular Biology **256**, 623 (1996).

[137] I. Suwan, H. Hussein, A. Hussein, and M. Daragmeh, Journal of Physics: Conference Series **869**, 012054 (2017).

[138] M. R. Shirts and J. D. Chodera, The Journal of Chemical Physics **129**, 124105 (2008).

[139] A. Brunet, C. Tardin, L. Salomé, P. Rousseau, N. Destainville, and M. Manghi, Macromolecules **48**, 3641 (2015).

[140] E. S. Sobel and J. A. Harpst, Biopolymers **31**, 1559 (1991).

[141] S. Geggier and A. Vologodskii, Proceedings of the National Academy of Sciences **107**, 15421 (2010).

[142] J. S. Mitchell, J. Glowacki, A. E. Grandchamp, R. S. Manning, and J. H. Maddocks, Journal of Chemical Theory and Computation **13**, 1539 (2017), pMID: 28029797.

[143] G. Lamour, J. B. Kirkegaard, H. Li, T. P. Knowles, and J. Gsponer, Source Code for Biology and Medicine **9**, 16 (2014).

[144] P. Cifra, Polymer **45**, 5995 (2004).

[145] H.-P. Hsu, W. Paul, and K. Binder, Macromolecules **43**, 3094 (2010).

[146] T. Schlick, B. Li, and W. Olson, Biophysical Journal **67**, 2146 (1994).

[147] A. Noy, T. Sutthibutpong, and S. A. Harris, Biophysical Reviews **8**, 233 (2016).

[148] J. S. Mitchell and S. A. Harris, Phys. Rev. Lett. **110**, 148105 (2013).

[149] C. W. Akey and K. Luger, Current Opinion in Structural Biology **13**, 6 (2003).

[150] T. Yadav and I. Whitehouse, Cell Reports **15**, 715 (2016).

[151] P. N. Dyer, R. S. Edayathumangalam, C. L. White, Y. Bao, S. Chakravarthy, U. M. Muthurajan, and K. Luger, in *Chromatin and Chromatin Remodeling Enzymes, Part A*, Methods in Enzymology, Vol. 375 (Academic Press, 2003) pp. 23–44.

[152] J. J. Hayes and K.-M. Lee, Methods **12**, 2 (1997).

[153] P. Pfaffle and V. Jackson, Journal of Biological Chemistry **265**, 16821 (1990).

[154] A. Bancaud, G. Wagner, N. e Silva, C. Lavelle, H. Wong, J. Mozziconacci, M. Barbi, A. Sivolob, E. Le Cam, L. Mouawad, J. L. Viovy, J. M. Victor, A. Prunell, N. Conde e Silva, C. Lavelle, H. Wong, J. Mozziconacci, M. Barbi, A. Sivolob, E. Le Cam, L. Mouawad, J. L. Viovy, J. M. Victor, and A. Prunell, Molecular Cell **27**, 135 (2007).

[155] C. Lavelle, P. Recouvreux, H. Wong, A. Bancaud, J. L. Viovy, A. Prunell, and J. M. Victor, Cell **139**, 1216 (2009).

[156] T. Furuyama and S. Henikoff, Cell **138**, 104 (2009).

[157] B. Brower-Toland, D. A. Wacker, R. M. Fulbright, J. T. Lis, W. L. Kraus, and M. D. Wang, Journal of Molecular Biology **346**, 135 (2005).

[158] S. Mihardja, A. J. Spakowitz, Y. Zhang, and C. Bustamante, Proceedings of the National Academy of Sciences of the United States of America **103**, 15871 (2006).

[159] M. Kruithof and J. Van Noort, Biophysical Journal **96**, 3708 (2009).

[160] H. Meng, K. Andresen, and J. Van Noort, Nucleic Acids Research **43**, 3578 (2015).

[161] M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, and H. E. Gaub, Science **276**, 1109 (1997).

[162] G. Fiorin, M. L. Klein, and J. Hénin, Molecular Physics **111**, 3345 (2013).

[163] R. A. Forties, J. A. North, S. Javaid, O. P. Tabbaa, R. Fishel, M. G. Poirier, and R. Bundschuh, Nucleic Acids Research **39**, 8306 (2011).

[164] F. T. Chien and T. Van Der Heijden, Biophysical Journal **107**, 373 (2014).

[165] D. Spakman, G. A. King, E. J. G. Peterman, and G. J. L. Wuite, Scientific Reports **10**, 9903 (2020).

[166] J. A. North, J. C. Shimko, S. Javaid, A. M. Mooney, M. A. Shoffner, S. D. Rose, R. Bundschuh, R. Fishel, J. J. Ottesen, and M. G. Poirier, Nucleic Acids Research **40**, 10215 (2012).

[167] W. Li, P. Chen, J. Yu, L. Dong, D. Liang, J. Feng, J. Yan, P. Y. Wang, Q. Li, Z. Zhang, M. Li, and G. Li, Molecular Cell **64**, 120 (2016).

[168] M. L. Bennink, S. H. Leuba, G. H. Leno, J. Zlatanova, B. G. De Grooth, and J. Greve, Nature Structural Biology **8**, 606 (2001).

[169] K. Struhl and E. Segal, Nat Struct Mol Biol **20**, 267 (2013).

[170] G. Meersseman, S. Pennings, and E. Bradbury, The EMBO Journal **11**, 2951 (1992).

[171] A. Flaus and T. Owen-Hughes, Molecular and Cellular Biology **23**, 7767 (2003).

[172] T. Sakaue, K. Yoshikawa, S. H. Yoshimura, and K. Takeyasu, Phys. Rev. Lett. **87**, 078105 (2001).

[173] I. M. Kulić and H. Schiessel, Phys. Rev. Lett. **91**, 148103 (2003).

[174] G. B. Brandani, T. Niina, C. Tan, and S. Takada, Nucleic Acids Research **46**, 2788 (2018).

[175] J. Lequieu, D. C. Schwartz, and J. J. de Pablo, Proceedings of the National Academy of Sciences **114**, E9197 (2017).

[176] H. Schiessel, J. Widom, R. F. Bruinsma, and W. M. Gelbart, Phys. Rev. Lett. **86**, 4414 (2001).

[177] I. Kulić and H. Schiessel, Biophysical Journal **84**, 3197 (2003).

[178] G. Pranami and M. H. Lamm, Journal of Chemical Theory and Computation **11**, 4586 (2015).

[179] J. T. Bullerjahn, S. von Bülow, and G. Hummer, The Journal of Chemical Physics **153**, 024116 (2020).

[180] T. Niina, G. B. Brandani, C. Tan, and S. Takada, PLOS Computational Biology **13**, 1 (2017).

[181] J. M. GOTTESFELD and D. A. MELTON, Nature **273**, 317 (1978).

[182] P. J. Fleming and K. G. Fleming, Biophysical Journal **114**, 856 (2018).

[183] S. J. Correll, M. H. Schubert, and S. A. Grigoryev, EMBO Journal **31**, 2416 (2012).

[184] S. Wei, S. J. Falk, B. E. Black, and T. H. Lee, Nucleic Acids Research **43** (2015), 10.1093/nar/gkv549.

[185] P. J. Robinson, W. An, A. Routh, F. Martino, L. Chapman, R. G. Roeder, and D. Rhodes, Journal of Molecular Biology **381**, 816 (2008).

[186] B. M. Turner, BioEssays **22**, 836 (2000).

[187] R. T. Simpson, Biochemistry **17**, 5524 (1978).

[188] D. V. Fyodorov, B.-R. Zhou, A. I. Skoultchi, and Y. Bai, Nature Reviews Molecular Cell Biology **19**, 192 (2018).

[189] J. Bednar, R. A. Horowitz, S. A. Grigoryev, L. M. Carruthers, J. C. Hansen, A. J. Koster, and C. L. Woodcock, Proceedings of the National Academy of Sciences **95**, 14173 (1998).

[190] A. Routh, S. Sandin, and D. Rhodes, Proceedings of the National Academy of Sciences **105**, 8872 (2008).

[191] A. L. Turner, M. Watson, O. G. Wilkins, L. Cato, A. Travers, J. O. Thomas, and K. Stott, Proceedings of the National Academy of Sciences **115**, 11964 (2018).

[192] A. Borgia, M. B. Borgia, K. Bugge, V. M. Kissling, P. O. Heidarsson, C. B. Fernandes, A. Sottini, A. Soranno, K. J. Buholzer, D. Nettels, B. B. Kragelund, R. B. Best, and B. Schuler, Nature **555**, 61 (2018).

[193] E. B. Gibbs and R. W. Kriwacki, Proceedings of the National Academy of Sciences **115**, 11868 (2018).

[194] F. L. Markley, Y. Cheng, J. L. Crassidis, and Y. Oshman, Journal of Guidance, Control, and Dynamics **30**, 1193 (2007).

[195] J. G. Gay and B. J. Berne, The Journal of Chemical Physics **74**, 3316 (1981).

[196] I. R. Cooke, K. Kremer, and M. Deserno, Phys. Rev. E **72**, 011506 (2005).

[197] W. M. Brown, M. K. Petersen, S. J. Plimpton, and G. S. Grest, The Journal of Chemical Physics **130**, 044901 (2009).

[198] M. P. Allen and G. Germano, Molecular Physics **104**, 3225 (2006).

[199] G. L. Lukacs, D. Lechardeur, P. Haggie, O. Seksek, N. Freedman, and A. Verkman, Journal of Biological Chemistry **275**, 1625 (2000).

[200] P. Li, S. Banjade, H.-C. Cheng, S. Kim, B. Chen, L. Guo, M. Llaguno, J. V. Hollingsworth, D. S. King, S. F. Banani, P. S. Russo, Q.-X. Jiang, B. T. Nixon, and M. K. Rosen, Nature **483**, 336 (2012).

[201] J. A. Ditlev, L. B. Case, and M. K. Rosen, Journal of Molecular Biology **430**, 4666 (2018).

[202] S. F. Banani, A. M. Rice, W. B. Peeples, Y. Lin, S. Jain, R. Parker, and M. K. Rosen, Cell **166**, 651 (2016).

[203] E. W. Martin, A. S. Holehouse, I. Peran, M. Farag, J. J. Incicco, A. Bremer, C. R. Grace, A. Soranno, R. V. Pappu, and T. Mittag, Science **367**, 694 (2020).

[204] J. R. Espinosa, J. A. Joseph, I. Sanchez-Burgos, A. Garaizar, D. Frenkel, and R. Collepardo-Guevara, Proceedings of the National Academy of Sciences of the United States of America **117**, 13238 (2020).

[205] P. Anderson and N. Kedersha, Nature Reviews Molecular Cell Biology **10**, 430 (2009).

[206] J. R. Espinosa, E. Sanz, C. Valeriani, and C. Vega, Journal of Chemical Physics **139**, 144502 (2013).

[207] R. García Fernández, J. L. F. Abascal, and C. Vega, Journal of Chemical Physics **124**, 144506 (2006).

[208] F. J. Blas, L. G. MacDowell, E. de Miguel, and G. Jackson, The Journal of Chemical Physics **129**, 144703 (2008).

[209] A. Ladd and L. Woodcock, Chemical Physics Letters **51**, 155 (1977).

[210] J. R. Espinosa, A. Garaizar, C. Vega, D. Frenkel, and R. Collepardo-Guevara, Journal of Chemical Physics **150**, 224510 (2019).

[211] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, The Journal of Chemical Physics **81**, 3684 (1984).

[212] V. B. Zhurkin and D. Norouzi, Biophysical Journal **120**, 577 (2021), publisher: Elsevier.

[213] X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren, and A. E. Mark, Angewandte Chemie International Edition **38**, 236 (1999).

[214] A. Sandoval-Pérez, R. M. L. Berger, A. Garaizar, S. E. Farr, M. A. Brehm, G. König, S. W. Schneider, R. Collepardo-Guevara, V. Huck, J. Rädler, and C. Aponte-Santamaría, Nucleic Acids Research **48**, 7333 (2020).

[215] C. A. Brackley, A. N. Morozov, and D. Marenduzzo, The Journal of Chemical Physics **140**, 135103 (2014).

[216] Journal of Molecular Biology **4**, 10 (1962).

[217] G. S. Manning, The Journal of Chemical Physics **51**, 924 (1969).

[218] D. Chakraborty, N. Hori, and D. Thirumalai, Journal of Chemical Theory and Computation **14**, 3763 (2018), pMID: 29870236.

[219] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, ACM Trans. Math. Softw. **22**, 469 (1996).

[220] H. Zhao, C. A. Brautigam, R. Ghirlando, and P. Schuck, Current Protocols in Protein Science **71**, 20.12.1 (2013).

[221] T. F. Miller, M. Eleftheriou, P. Pattnaik, A. Ndirango, D. Newns, and G. J. Martyna, The Journal of Chemical Physics **116**, 8649 (2002).

[222] R.-H. Huang, D. H. Fremont, J. L. Diener, R. G. Schaub, and J. E. Sadler, Structure **17**, 1476 (2009).

Appendix A

Algorithms and analysis

A.1 Quaternions and Rotations

A unit quaternion $\underline{q}$,

$$\underline{q} = (q_w, q_x, q_y, q_z) = q_w + q_x\mathbf{i} + q_y\mathbf{j} + q_z\mathbf{k}, \tag{A.1}$$

$$|\underline{q}| = \sqrt{(q_w^2 + q_x^2 + q_y^2 + q_z^2)} = 1,$$

describes a rotation of angle θ about axis $\mathbf{a}$, where,

$$\underline{q} = (\cos(\theta/2), a_x \sin(\theta/2), a_y \sin(\theta/2), a_z \sin(\theta/2)). \tag{A.2}$$

This is equivalent to an orthogonal rotation matrix $\mathbf{R}$:

$$\mathbf{R} = \begin{pmatrix} q_w^2 + q_x^2 - q_y^2 - q_z^2 & 2(q_x q_y - q_w q_z) & 2(q_x q_z + q_w q_y) \\ 2(q_x q_y + q_w q_z) & q_w^2 - q_x^2 + q_y^2 - q_z^2 & 2(q_y q_z - q_w q_x) \\ 2(q_x q_z - q_w q_y) & 2(q_y q_z + q_w q_x) & q_w^2 - q_x^2 - q_y^2 + q_z^2 \end{pmatrix}. \tag{A.3}$$

The columns of $\mathbf{R}$ are orthogonal unit vectors

$$\mathbf{R} = (\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}})\,. \tag{A.4}$$

These three representations of a rotation are equivalent and we convert between them when a certain representation is more useful.

$$\underline{q} \equiv \mathbf{R} \equiv (\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}})\,. \tag{A.5}$$

In the same way that a position vector gives the position of a point relative to the origin, a rotation gives the orientation of a body relative to the simulation frame, which is always assumed to have orientation

$$\underline{q} = (1, 0, 0, 0),\ \mathbf{R} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{A.6}$$

To convert from $\mathbf{R}$ to $\underline{q}$ the following can be used:

$$\begin{aligned} q_w &= \sqrt{1 + \mathrm{Trace}(\mathbf{R})}/2, \\ q_x &= (R_{zy} - R_{yz})/(4q_w), \\ q_y &= (R_{xz} - R_{zx})/(4q_w), \\ q_z &= (R_{yx} - R_{xy})/(4q_w). \end{aligned} \tag{A.7}$$

However, in practice care must be taken to avoid divide by zero errors when the trace is negative. Therefore we use the follow procedure replicated from LAMMPS [94] source code:

```
def R_to_q(R):
    """ Python code translated from LAMMPS c++"""

    ex=R[:,0]
    ey=R[:,1]
    ez=R[:,2]

    # squares of quaternion components
    q0sq = 0.25 * (ex[0] + ey[1] + ez[2] + 1.0)
    q1sq = q0sq - 0.5 * (ey[1] + ez[2])
    q2sq = q0sq - 0.5 * (ex[0] + ez[2])
    q3sq = q0sq - 0.5 * (ex[0] + ey[1])

    q = np.array([0.0, 0.0, 0.0, 0.0])
    # some component must be greater than 1/4 since they sum to 1
    # compute other components from it

    if q0sq >= 0.25:
        q[0] = np.sqrt(q0sq)
        q[1] = (ey[2] - ez[1]) / (4.0 * q[0])
        q[2] = (ez[0] - ex[2]) / (4.0 * q[0])
        q[3] = (ex[1] - ey[0]) / (4.0 * q[0])
    elif q1sq >= 0.25:
        q[1] = np.sqrt(q1sq)
        q[0] = (ey[2] - ez[1]) / (4.0 * q[1])
        q[2] = (ey[0] + ex[1]) / (4.0 * q[1])
        q[3] = (ex[2] + ez[0]) / (4.0 * q[1])
    elif q2sq >= 0.25:
        q[2] = np.sqrt(q2sq)
        q[0] = (ez[0] - ex[2]) / (4.0 * q[2])
        q[1] = (ey[0] + ex[1]) / (4.0 * q[2])
        q[3] = (ez[1] + ey[2]) / (4.0 * q[2])
    elif q3sq >= 0.25:
        q[3] = np.sqrt(q3sq)
        q[0] = (ex[1] - ey[0]) / (4.0 * q[3])
        q[1] = (ez[0] + ex[2]) / (4.0 * q[3])
        q[2] = (ez[1] + ey[2]) / (4.0 * q[3])

    return q
```

To rotate a vector $\mathbf{v}$ about an axis $\hat{\mathbf{a}}$ by angle θ we can also use the Rodrigues rotation formula,

$$\mathbf{v}' = \mathbf{v}\cos(\theta) + (\hat{\mathbf{a}} \times \mathbf{v})\sin(\theta) + \hat{\mathbf{a}}(\hat{\mathbf{a}} \cdot \mathbf{v})(1 - \cos(\theta)). \tag{A.8}$$

The quaternion product $\underline{r} = \underline{a}\,\underline{b}$ is given by:

$$\begin{aligned} r_w &= a_w b_w - a_x b_x - a_y b_y - a_z b_z, \\ r_x &= a_w b_x + b_w a_x + a_y b_z - a_z b_y, \\ r_y &= a_w b_y + b_w a_y + a_z b_x - a_x b_z, \\ r_z &= a_w b_z + b_w a_z + a_x b_y - a_y b_x. \end{aligned} \tag{A.9}$$

The inverse of a unit quaternion is given by:

$$\underline{q}^{-1} = (q_w, -q_x, -q_y - q_z). \tag{A.10}$$

A.2 Algorithm for calculation of helical parameters

To calculate the helical parameters between two base-pairs we use the SCHNAAP procedure [130]. The two base-pairs (DNA Ellipsoids) have positions and orientation quaternions $\mathbf{r}_1, \underline{q}_1$ and $\mathbf{r}_2, \underline{q}_2$ respectively.

1. Convert quaternions $\underline{q}_1$ and $\underline{q}_2$ to matrices $\mathbf{T}_1$ and $\mathbf{T}_2$ whose columns are the $\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}$ axis direction vectors.

2. Calculate the roll-tilt angle Γ:

$$\Gamma = \cos^{-1}(\mathbf{z}_1 \cdot \mathbf{z}_2). \tag{A.11}$$

3. Calculate the roll-tilt axis $\mathbf{rt}$:

$$\mathbf{rt} = \hat{\mathbf{z}}_1 \times \hat{\mathbf{z}}_2. \tag{A.12}$$

4. Rotate base-pair 1 and 2 about $\mathbf{rt}$ by $+\Gamma/2$ and $-\Gamma/2$ respectively:

$$\mathbf{T}'_1 = \mathbf{R}(\mathbf{rt}, +\Gamma/2)\mathbf{T}_1, \tag{A.13}$$
$$\mathbf{T}'_2 = \mathbf{R}(\mathbf{rt}, -\Gamma/2)\mathbf{T}_2, \tag{A.14}$$

   where $\mathbf{R}(\mathbf{a}, \theta)$ is an orthogonal matrix that describes a rotation of θ about axis $\mathbf{a}$.

5. The mid-step matrix $\mathbf{T}_{\text{ms}}$ is the mean of the rotated matrices:

$$\mathbf{T}_{\text{ms}} = \frac{1}{2}(\mathbf{T}'_1 + \mathbf{T}'_2). \tag{A.15}$$

6. Twist Ω is the angle between the transformed y-axis:

$$\Omega = \cos^{-1}(\hat{\mathbf{y}}'_1 \cdot \hat{\mathbf{y}}'_2) \tag{A.16}$$

7. Calculate ϕ the angle between the roll-tilt axis and the mid-step y-axis:

$$\phi = \cos^{-1}(\hat{\mathbf{rt}} \cdot \hat{\mathbf{y}}_{\text{ms}}). \tag{A.17}$$

8. Roll ρ and tilt τ are given by:

$$\rho = \Gamma \cos(\phi), \tag{A.18}$$
$$\tau = \Gamma \sin(\phi). \tag{A.19}$$

9. Shift D_x, slide D_y, and rise D_z are calculated as:

$$(D_x, D_y, D_z)^\intercal = T(\mathbf{r}_2 - \mathbf{r}_1). \tag{A.20}$$

There is an edge case in the algorithm if the two base-pairs have the same $\hat{\mathbf{z}}$: Γ in step 2 becomes zero and the cross product in step 3 results in a undefined roll-tilt axis. We deal with this as follows: if $\hat{\mathbf{z}}_1 \cdot \hat{\mathbf{z}}_2 == 1$ then

1. Set roll and tilt to zero

$$\rho = 0, \tag{A.21}$$

$$\tau = 0. \tag{A.22}$$

2. Twist Ω is the angle between the y axis:

$$\Omega = \cos^{-1}(\hat{\mathbf{y}}_1 \cdot \hat{\mathbf{y}}_2), \tag{A.23}$$

3. The mid-step matrix is the the mean of the two orientation matrices

$$\mathbf{T}_{\text{ms}} = \frac{1}{2}(\mathbf{T}_1 + \mathbf{T}_2). \tag{A.24}$$

4. Shift D_x, slide D_y, and rise D_z are calculated as:

$$(D_x, D_y, D_z)^\intercal = \mathbf{T}(\mathbf{r}_2 - \mathbf{r}_1). \tag{A.25}$$

A.3 Calculation of Sedimentation coefficient using the HullRad method

To calculate sedimentation coefficients for the chemically specific model we use the HullRad method [182]. This uses a convex hull model to estimate the hydrodynamic volume of the molecule. For completeness the method will be described here. A convex hull is mathematically defined as the smallest convex envelope that contains a set of points. These points are the 3D coordinates of our molecular structure. The Python SciPy library method ConvexHull is used to compute the convex hull. This method use the Qhull library [219]. From the convex hull we get the volume V_H and the surface area A_H. We then calculate the hydration shell thickness $V_W = A_H * 2.8$ and add this to the hull volume to get the total volume $V_H W = V_H + V_W$. Additionally, we find the maximum distance between vertices in the convex hull D_{max} and use this to get the axial ratios, $a = (D_{max}/2.0)$, $b = \sqrt{(3V_H)/(4\pi a)}$. These are used to compute the translational shape factor

$$F_t = \sqrt{\frac{\sqrt{1-(b/a)^2}}{(b/a)^{(2/3)} \log\left[(1+\sqrt{1-(b/a)^2})/(b/a)\right]}}, \tag{A.26}$$

We then calculate the translational hydrodynamic radius

$$R_T = F_t \sqrt[3]{\frac{3V_{HW}}{4\pi}}, \tag{A.27}$$

The Sedimentation coefficient is then calculated as

$$s = 10^8 \left(\frac{M - M\bar{v}\rho_{20,w}}{N_A 6\pi\eta_0 R_T} \right), \tag{A.28}$$

where M is the total molar mass, $\bar{v}$ is the total partial specific volume, $\rho_{20,w}$ is the density of water at 20C, η_0 is viscosity of water at 20C and,

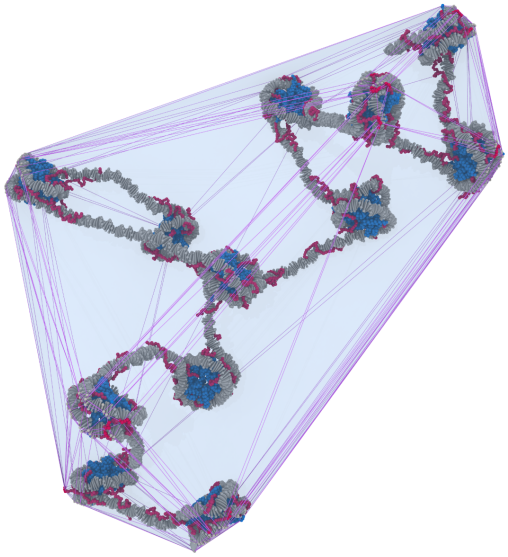

Figure A.1: **Convex hull of a chromatin fiber, used to compute the sedimentation coefficient.**

$$M = \sum_i m_i, \tag{A.29}$$

$$\bar{v} = \frac{1}{M} \sum_i m_i v_i, \tag{A.30}$$

where m_i and v_i are the molar masses and specific volumes of the individual beads, listed in table A.1

A.3.1 Rationale for using sedimentation coefficients

The sedimentation coefficient is a property of molecules in solution that can be measured by ultra-centrifuge experiments [220]. It defined in terms of the sedimentation velocity and the applied acceleration,

$$S = v_s/a, \tag{A.31}$$

where v_s, the sedimentation velocity, is the linear speed of movement of the sedimentation boundary in the centrifuge and a is the acceleration applied by the centrifuge, $a = \omega^2 r$, where ω is the angular velocity of centrifuge and r is the distance from the center. Sedimentation coefficient has units of time, expressed in Svedbergs, symbol S, where $1\text{S} = 10^{-13}$ seconds. The sedimentation coefficient is also related the molecular properties by [220]

$$S = \frac{M(1 - \bar{v}\rho)}{Nf}, \tag{A.32}$$

where M is the molecular weight, $\bar{v}$ is the partial specific volume, ρ is the solution density, f is the molecules friction coefficient, and N is Avogadro's number. This demonstrates the sedimentation coefficient is completely defied by the mass and shape of the molecule, and the properties of the solvent. Thus we can approximate

Bead type	m (g/mol)	$\bar{v}$ (mL/g)
ALA	71.08	0.74
ARG	156.20	0.73
ASN	114.10	0.63
ASP	115.10	0.60
CYS	103.10	0.62
GLN	128.10	0.68
GLU	129.10	0.66
GLY	57.05	0.64
HIS	137.10	0.67
ILE	113.20	0.90
LEU	113.20	0.90
LYS	128.20	0.82
MET	131.20	0.75
PHE	147.20	0.77
PRO	97.12	0.76
SER	87.08	0.63
THR	101.10	0.70
TRP	186.20	0.74
TYR	163.20	0.71
VAL	99.07	0.86
DNA ellipsoid	500	0.65
DNA phosphate	62.97	0.501

Table A.1: Molar masses m and partial specific volumes $\bar{v}$, of CG beads used in the HullRad method calculation of sedimentation coefficients.

the calculation from our simulations, either by the HullRad method described above, or by the simpler method described in 6.3.1, and compare with the sedimentation coefficients reported from experiments. The sedimentation coefficient is approximately inversely proportional to the radius of gyration, $S \sim 1/R_g$, i.e a smaller R_g correlates with a larger S.

A.4 Amount of unwrapped DNA

For the simulations of breathing chromatin it becomes slightly difficult to define which DNA beads are nucleosomal and which are linker. This is because at the higher salt values the dense chromatin structures have DNA in contact with the nucleosome core proteins that is not part of that nucleosome, so simply computing the protein-DNA contacts will not work. To overcome this we developed the following procedure: first we record which protein beads are bonded to the DNA in simulations of non-breathing nucleosomes, these protein beads are located circularly around the nucleosome in the locations where the DNA is typically wrapped. Then, for each frame in the breathing trajectory, we compute the contacts between the DNA and the aforementioned protein beads. For each nucleosome we now have a list of bound DNA beads. We then compute the median DNA bead in terms of index along the DNA sequence. This is approximately the center bead of that nucleosome's nucleosomal DNA. We then look forwards and backwards along the DNA sequence, within the range of maximum and minimum indices of the bound DNA beads, and unless a large continuous region of unbound DNA (>100bp) is found, all the DNA between the maximum and minimum limits is added. Each nucleosome now has a contiguous section of nucleosomal DNA assigned to it. Finally the list of nucleosomal DNA is checked for overlaps and any are removed to ensure that each DNA bead can only be a member of one nucleosome. The average amount of unwrapped DNA per nucleosome is then computed as:

$$N_{\text{unwrapped}} = (147N_n - N_{\text{nucleosomal DNA}})/N_n, \tag{A.33}$$

where N_n is the number of nucleosomes, $N_{\text{nucleosomal DNA}}$ is the total number of nucleosomal DNA beads, and 147 is the typical number of base pairs of DNA wrapped round one nucleosome.

A.5 Inter-nucleosome interactions

The relative orientation of two nucleosomes can be categorised into three states: face-face (ff), face-side (fs), and side-side (ss) as illustrated in Figure A.2. To characterize these, we compute the nucleosome orientation matrices—the columns of which are the orthogonal unit axis vectors of the nucleosome. The center of a nucleosome is defined as the center of mass of the globular domain beads. The x axis passes through the nucleosome dyad and the z axis points perpendicularly out of the nucleosome "face" as shown in Figure A.2.

Below, we explain the procedure used to categorize the relative orientation of

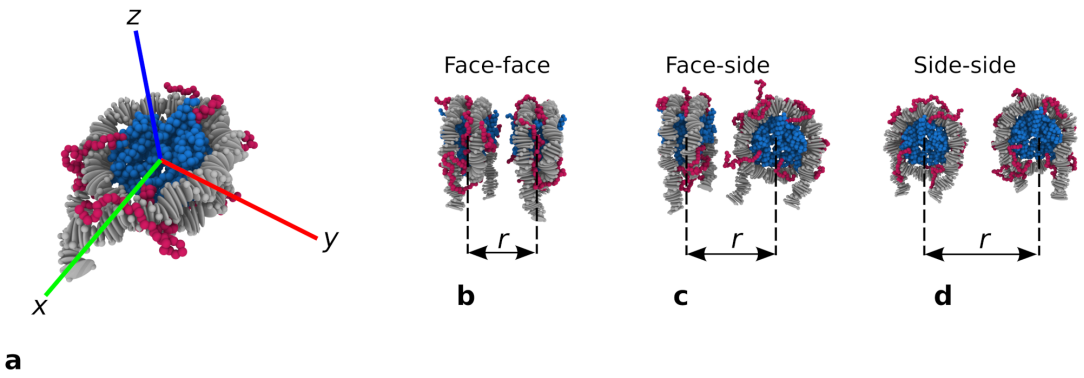

Figure A.2: **Definition of nucleosome pair orientations. (a)** Nucleosome orientation axis: x points from the center of the nucleosome to the dyad position, z points out of the top face, and $y = z \times x$. **(b-d)** Nucleosome–nucleosome interaction configurations, r is the center-to-center distance between two nucleosomes.

the nucleosomes. We define the angles $\{\alpha, \beta_i, \beta_j\}$ as follows:

$$\hat{\mathbf{z}}_i \cdot \hat{\mathbf{z}}_j = \cos\alpha, \tag{A.34}$$

$$\hat{\mathbf{z}}_i \cdot \hat{\mathbf{r}} = \cos\beta_i, \tag{A.35}$$

$$\hat{\mathbf{z}}_j \cdot \hat{\mathbf{r}} = \cos\beta_j, \tag{A.36}$$

where $\hat{\boldsymbol{r}}$ is the unit vector pointing from the center of the i^{th} nucleosome to the center of the j^{th} nucleosome, and $\hat{\boldsymbol{z}}_i$ and $\hat{\boldsymbol{z}}_j$ are the unit z-axis vectors of i^{th} and j^{th} nucleosomes respectively. We then use the following algorithm:

```
if α < 45° or α > 135°:
    if β_i < 45° or β_i > 135° or β_j < 45° or β_j > 135°:
        Face-face
    else:
        Side-side
else:
    Face-side
```

We then construct three interaction matrices M_{ij}^{μ} between the i^{th} and j^{th} nucleosomes, one for each relative orientation $\mu = \{\text{ff}, \text{fs}, \text{ss}\}$:

$$M_{ij}^{\mu} = \frac{1}{N_t} \sum_t C_{ij}^{\mu}(t), \tag{A.37}$$

$$C_{ij}^{\mu}(t) = \begin{cases} 1, & \text{if nucleosomes } i \text{ and } j \text{ are in contact, and have a type } \mu \text{ relative orientation,} \\ 0, & \text{otherwise,} \end{cases} \tag{A.38}$$

where t is the timestep. The sum is taken over all N_t snapshots used in the analysis. Two nucleosomes are defined to be in "contact" when the center to center distance between them is $< 110\,\text{Å}$. The interaction matrices can then be projected onto a 1D map to describe the relative intensity of interactions between nucleosomes separated by $(k-1)$ neighbours,

$$I^{\mu}(k) = \frac{1}{N_n} \sum_i M_{i,i+k}^{\mu}. \tag{A.39}$$

This sum is equivalent to taking the means of the diagonals of M_{ij}^{μ}.

A.6 Molecular-level inter-nucleosome contacts

To calculate the molecular-level inter-nucleosome contacts we use a similar procedure to the inter-nucleosome interactions but at bead level rather than nucleosome level. For each frame in the simulation trajectory the total contact matrix is computed for all beads.

$$M_{ij} = \begin{cases} 1, & \text{if beads } i \text{ and } j \text{ are in contact,} \\ 0, & \text{otherwise,} \end{cases} \tag{A.40}$$

where "in contact" here is true when the distance between beads i and j is less than $((\sigma_i + \sigma_j)/2 + 1\text{Å})$. For each bead three sums are performed—one each counting up the contacts with: DNA, Histone tail, and Globular domain beads. Importantly the contacts are only counted if they are located in different nucleosomes. For the breathing DNA this is non-trivial. To proceed we defined nucleosomal DNA using the same method as in Section A.4. The remaining linker DNA is then assigned to the nucleosome it is closest to (in term of DNA sequence, not spatial distance). This process enables the contacts to be computed for interactions between different nucleosomes that would otherwise be dominated by the intra-nucleosome contacts.

$$C^{\mathrm{x}}(i) = \sum_{j \text{ in x and in different nucleosome to } i} M_{ij}, \ \mathrm{x} = \{\text{DNA, Histone tails, Globular domain}\}. \tag{A.41}$$

$C^{\mathrm{x}}(i)$ is a list with a length equal to the total number of beads. There are three of them, one for each x = DNA, histone tails, globular domains. $C^{\mathrm{x}}(i)$ is then averaged over all nucleosomes and all timesteps in the trajectory, and normalized by its maximum value C_{MAX}. To generate the visualizations of the nucleosome contacts each bead is given a RGB color according to:

$$\mathrm{color}_i = [\mathrm{red}, \mathrm{green}, \mathrm{blue}] = 255 \frac{\log_{10}\left(C_{\mathrm{MAX}} \times [C^{\mathrm{DNA}}(i), C^{\text{Globular domain}}(i), C^{\text{Histone tails}}(i)]\right)}{\log_{10}(C_{\mathrm{MAX}})}. \tag{A.42}$$

Where RBG values are integers in the range 0-255.

A.7 Radius of gyration

We compute the radius of gyration R_g as

$$R_g = \sqrt{\frac{1}{N}\left\langle \sum_i^N |\mathbf{r}_i - \mathbf{r}_{\mathrm{COM}}|^2 \right\rangle}, \tag{A.43}$$

$$\mathbf{r}_{\mathrm{COM}} = \frac{1}{N}\sum_i^N \mathbf{r}_i, \tag{A.44}$$

where $\mathbf{r}_i$ are the coordinates of particle i and N is the total number of particles in the molecule. The angular brackets indicate an average taken over all timesteps used in the analysis.

A.8 LAMMPS 'fix' algorithms

Here we list the algorithms and numerical methods corresponding to the various LAMMPS fix commands we use.

A.8.1 fix nve

`fix nve` is a velocity verlet integrator:

$$\begin{aligned} v^{i+1/2} &= v^i + \frac{dt}{2m}F^i, \\ x^{i+1} &= x^i + dtv^{i+1/2}, \\ F^{i+1} &= -\nabla E(x^{i+1}), \\ v^{i+1} &= v^{i+1/2} + \frac{dt}{2m}F^{i+1}. \end{aligned} \tag{A.45}$$

A.8.2 fix rigid/nve

`fix rigid/nve` and fix `rigid/nve/small` use the rigid body integrator of Miller et al [221].

A.8.3 fix langevin

`fix langevin` applies a Langevin thermostat by modifying the force calculation in `fix nve` to

$$F^{i+1} = -\nabla E(x^{i+1}) + F_{\text{friction}} + F_{\text{random}}, \tag{A.46}$$

$$F_{\text{friction}} = -\gamma v^{i+1/2}, \tag{A.47}$$

$$F_{\text{random}} = \sqrt{\frac{2\gamma k_B T}{dt}}R, \tag{A.48}$$

where $\gamma = m/t_{\text{damp}}$ and R is a normally distributed random number. When the GJF formulation is used F is instead calculated as

$$F^{i+1} = \frac{1}{1+\frac{\gamma dt}{2m}}\left[-\nabla E(x^{i+1}) - \gamma v^{i+1/2} + \sqrt{\frac{2\gamma k_B T}{dt}}\frac{(R^{i+1}+R^i)}{2}\right]. \tag{A.49}$$

When the angular momentum is included the torque acting on the finite size particles is modified in an analogous way to the force:

$$\tau^{i+1} = \tau_{\text{E}}^{i+1} + \tau_{\text{friction}} + \tau_{\text{random}}, \tag{A.50}$$

where τ_{E}^{i+1} is the torque due to the potential energy function E and

$$\tau_{\text{friction}} = -\frac{I}{t_{damp}}\omega, \tag{A.51}$$

$$\tau_{\text{random}} = \sqrt{\frac{I}{t_{\text{damp}}}\frac{2k_B T}{dt}}R, \tag{A.52}$$

where I is the inertia and ω is the angular velocity.

Appendix B

Additional applications

B.1 DNA binding to the blood protein von Willebrand factor (vWF)

This work was carried out as part of a collaboration with the group of Dr. Camilo Aponte-Santamaria and has been published [214].

The von Willebrand factor is a protein present in blood plasma that is involved in hemostasis (the process of preventing bleeding). Additionally present in blood plasma is extra-cellular DNA in the from of neutophil extracellular traps (NETs). NETs are macro-molecular meshes composed of DNA, protein, and antimicrobial agents that ensnare and kill pathogens as part of the immune response. The misregulation of NETs is linked to pathological conditions, however the process by which NETs adheres to blood vessels is not completely understood. vWF is likely to play a role in this process, specifically the A1 domain of vWF is thought to bind directly to DNA due to the presence of a positively charged region.

Angélica Sandoval-Pérez and Camilo Aponte-Santamaría carried out atomistic simulations of the vWF A1 domain binding to DNA which indicated that electrostatics are responsible for vWF to DNA binding. We supported the research by performing CG simulations of the same system.

B.1.1 Methods

The conformation of the vWF A1 domain was taken from the X-ray crystallography structure PDB identification code 3HXO [222]. Following our previous method in section 3.3 we place CG beads on the C_α atoms and force the preservation of the shape and structure by constructing a Gaussian elastic Network Model with a cutoff distance of 7 Å and a bond strength of 10 kcal Å^{-2} mol^{-1}. The GNM uses harmonic bonds of the form

$$V_{\text{Prot. bonds}} = \frac{1}{2}k(r - r_0)^2, \tag{B.1}$$

which was implemented in LAMMPS using `bond_style harmonic`.

The DNA was constructed in a relaxed linear configuration with a length of 42 bp using our base-pair resolution DNA model described in section 3.2. We used two sequences: ARC1172 (GGC GTG CAG TGC CTT CGG CCG TGC GGT GCC TCC GTC ACG CCT), and PolyAT (AAA AAA AAA AAA AAA AAA AAA

AAA AAA AAA AAA AAA AAA AAA). The DNA bonds use the rigid base-pair potential.

Because the protein is globular (i.e all beads are part of the GNM) there are no intra-protein pairwise terms. The charged protein beads and DNA phosphate sites interact via a Debye-Hückel potential which approximates screening by counterions in solution:

$$V_{\text{Electrostatic}} = \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r} e^{-r/\lambda_d}, \tag{B.2}$$

where q_i and q_j are the charges (-1 for DNA phosphates and see table 3.2 for protein beads), ϵ is the vacuum permitivity, ϵ_r is the relative permitivity (set to 80 for water), and λ_d is the Debye screening length which is set to 8 Å corresponding to 0.15 M monovalent salt concentration. Furthermore, we include a protein-DNA excluded volume term in the form of a shifted-truncated Lennard-Jones potential

$$V_{\text{LJ}} = \begin{cases} 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{2}\right] + \epsilon, & r \leq 2^{1/6}\sigma, \\ 0, & r > 2^{1/6}\sigma. \end{cases} \tag{B.3}$$

This differs slightly from the form used in the chromatin model as it is completely repulsive, we did this because the attractive part of the potential in the chromatin model was optimized to represent DNA-nucleosome interaction and we are not sure how transferable that part of the model is. Furthermore, we are interested in electrostatic effects in this study so by only including pairwise attraction from electrostatics we can investigate if the interactions are sufficient for vWF A1 to bind to DNA. We set $\epsilon = 0.1\,kcal/mol$, $\sigma = 4$ Å for protein–phosphate and $\epsilon = 0.01\,kcal/mol$, $\sigma = 8$ Å for protein–DNA base-pair.

We positioned the DNA strand and the protein in a cubic periodic simulation box with dimensions $400 \times 400 \times 400$ Å. This is large enough such that there is no self interaction between periodic images, i.e. the length of the DNA is approximately 138 Å. The DNA and protein are initially unbound and kept separated by a minimum distance of 20 Å. We performed 32 simulations for each DNA sequence using different random starting configurations. Each trajectory was run for 3.5 µs in the NVT ensemble using a Langevin thermostat at 300 K with a relaxation time of 100 ps and a timestep of 10 fs.

B.1.2 Results and Discussion

An example initial configuration in the unbound state is shown in figure B.1a. All trajectories were observed to transition to the bound state within approximately 1 ns. An example bound state is shown in figure B.1b. Observing the pairwise potential energy timeseries of the simulations we see occasional unbinding and rebinding events, an example time series is shown in figure B.1c, the step like jumps indicated by the red crosses show the unbound states. Averaging across all trajectories we see that approximately 99.9% of the time is spent in the bound state. The atomistic resolution simulations carried out in the referenced paper do not see unbinding after the initial association of vWF A1 to the DNA strand. The rapid association of vWF A1 to the DNA is observed in both our CG simulations and the atomistic simulations. Of the 84 repeats performed in the atomistic simulations 80% are observed to associate within 50 ns and 95% within 100 ns. Although rigorous direct comparison of these timescales between the atomistic and our CG simulations is

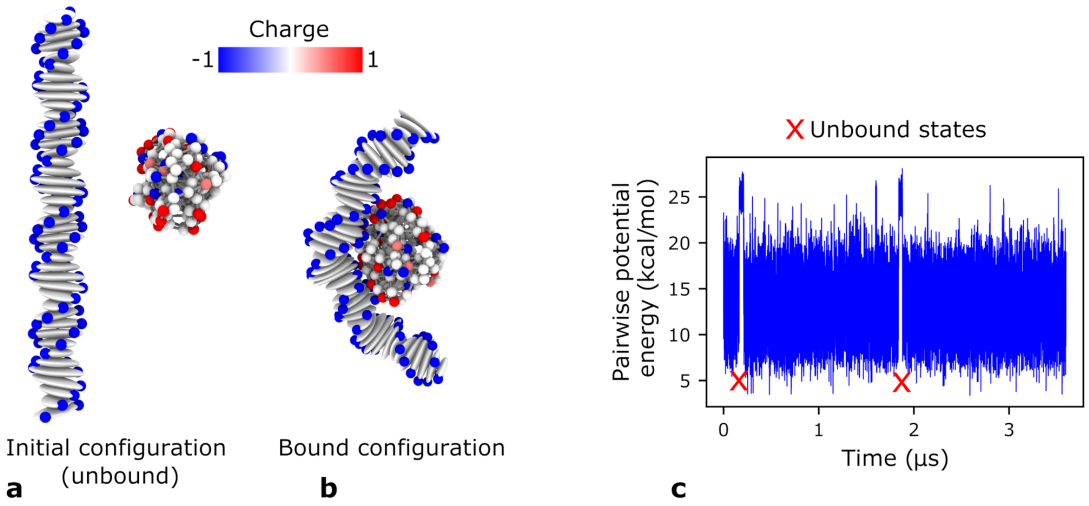

Figure B.1: **DNA binding to vWF**. **(a)** Shows the initial unbound configuration. **(b)** Shows a bound configuration. Both are color coded according to the charge. **(c)** Example timeseries of the pairwise potential energy of a simulation, the unbinding and rebinding events are marked by red crosses.

not possible (i.e. there are differences in box sizes, starting configurations, Langevin dynamics vs molecular dynamics etc) it suggests our CG simulation timescales are of the order of 10 to 100 times that of the atomistic simulation timescales. This vastly increased sampling could explain why we observe occasional unbinding events during the simulations while the atomistic ones do not.

Investigating the equilibrium properties of the bound DNA we measure the bending angle distribution. The bending angle, illustrated in B.2a, was defined as the angle between a vector from the center of the first base-pair and the center of the DNA molecule and a vector between the center point and the last base-pair. The center point is the mean position of the middle two base-pairs (the 21st and 22nd). We calculate that the mean angle for bound ARC DNA is 30 degrees with a standard deviation of 16 degrees, and for the unbound ARC DNA it is 25 degrees with an standard deviation of 12 degrees. For PolyAT bound it is 26 degrees with standard deviation of 14 and for PolyAT unbound it is 22 with standard deviation of 11 degrees. The distributions of these angles are plotted in figure B.2b. Note that the the unbound values were calculated from simulations with just DNA and no protein for the similar simulation lengths similar to the DNA plus protein simulations.

To investigate the mode of binding between DNA and vWF A1 domain we compute the contacts between the amino-acid beads and the DNA. A contact is defined when the distance between an amino-acid bead and a DNA base-pair bead is less than 9Å or the distance between an amino acid bead and a DNA phosphate bead is less than 5Å. A contact between an amino acid and a DNA-base pair is only counted once (there are 3 beads composing a base-pair). For each simulation frame we now have a list denoting if an amino acid bead is in contact with any DNA bead bead and vice-versa we have a list denoting if a base-pair is in contact with any amino-acid bead. We take the mean over all timesteps, which for each amino-acid bead gives the proportion of time it is in contact with DNA. Similarly for each DNA

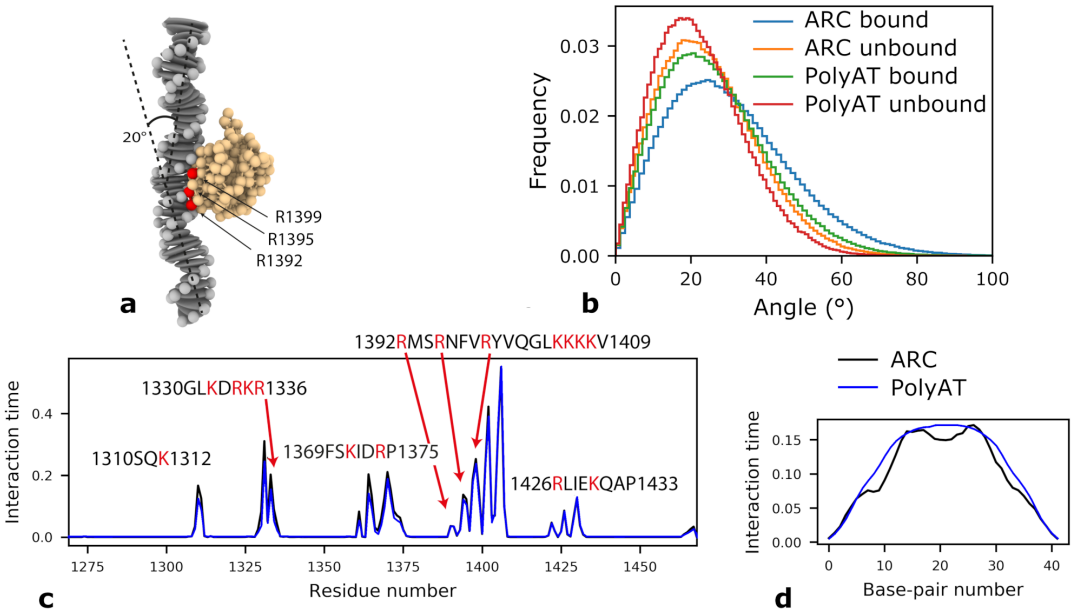

Figure B.2: **DNA binding to vWF**. **(a)** Shows an equilibrium simulation snapshot, the DNA bending angle is labelled. Three specific arginine amino-acids are indicated and colored red. **(b)** The distributions of the DNA angle for all 4 simulations types indicated in the legend. **(c)** Interactions plots (normalized contacts) for each amino-acid with DNA for the ARC (black line) and polyAT (blue line) DNA sequences. **(d)** Interaction plots (normalized contacts) for each DNA base-pair with the protein for the ARC (black line) and polyAT (blue line) DNA sequences.

base-pair it gives the proportion of time it is contact with the protein. This is what we term 'interaction time' in the plot in figures B.2c and d. The peaks correspond to the positively charged arginine (R) and lysine (K) amino-acids, which have been labelled in terms of their sequence, following the numbering scheme in [214]. These peaks are in good agreement with the atomistic simulations [214] demonstrating that vWF does bind to DNA predominately via electrostatic interactions.

Appendix C

Guide for software use

The source code and scripts needed to run the model presented in this thesis are provided in the GitHub repository https://github.com/CollepardoLab/CollepardoLab_Chromatin_Model and https://doi.org/10.6084/m9.figshare.13663685.v1.

System requirements

Linux with C++ compilers with MPI. Tested on CSD3 peta-4 cluster https://www.hpc.cam.ac.uk/systems/peta-4 with Intel 2017 compiler suite.

Installation guide

LAMMPS needs to be compiled with our custom code

1. clone a copy of LAMMPS

   ```
   git clone https://github.com/lammps/lammps.git
   ```

2. checkout stable version 3rd March 2020

   ```
   cd lammps
   git checkout tags/stable_3Mar2020 -b stable
   ```

3. copy all our code from `lammps_custom_code` into `lammps/src`

4. move `Makefile_DNA_mpi` from `lammps/src` into `lammps/src/MAKE`

5. Install required LAMMPS packages

   ```
   make yes-asphere
   make yes-rigid
   make yes-molecule
   ```

6. compile using our makefile, note this is for Intel compilers only

   ```
   make DNA_mpi
   ```

7. the executable will be `lmp_DNA_mpi`

Demo simulation

To run a single nucleosome system:

1. move to the "demo" directory
2. run with LAMMPS

   ```
   mpirun -np 1 ./lmp_DNA_mpi -in in.run
   ```

It will produce a LAMMPS trajectory file "dna.dump" this can be viewed in Ovito https://www.ovito.org/. Visible molecular dynamics will be observable after a few minutes run-time on a single core.

Instructions to reproduce select results

Running chemically-specific 12N chromatin HREMD simulations

1. The files are in `main_simulations/input_scripts/chemically_specific_12N_165NRL_HREMD`. The LAMMPS input scripts are `in.hremd_breathing` and `in.hremd_nonbreathing`
2. run LAMMPS using at least 16 cores

   ```
   mpirun -np 16 ./lmp_DNA_mpi -partition 16x1 -in in.hremd_breathing
   ```

3. These simulations for both breathing and non-breathing will generate the trajectories for the results in section 4.9.

Running minimal model coexistence simulations

1. The files are in `main_simulations/input_scripts/minimal_coexistence/`
2. run a coexistence simulation

   ```
   mpirun -np 16 ./lmp_DNA_mpi -in in.run
   ```

3. To reproduce the phase diagram (figure 6.5) one would need to vary the parameters E1 and A in the input script. These correspond to the variables E and S respectively in table 5.3. Additionally the breathing and non-breathing structures can be used by changing which data file is read in (`data_nonb.txt` or `data_b.txt`).

Appendix D

Supplementary tables

Atom type ID	Region	Atom type represented
1	Histone tails	ALA
2		ARG
3		ASN
4		ASP
5		CYS
6		GLN
7		GLU
8		GLY
9		HIS
10		ILE
11		LEU
12		LYS
13		MET
14		PHE
15		PRO
16		SER
17		THR
18		TRP
19		TYR
20		VAL
21	Histone globular domains	ALA
22		ARG
23		ASN
24		ASP
25		CYS
26		GLN
27		GLU
28		GLY
29		HIS
30		ILE
31		LEU
32		LYS
33		MET
34		PHE
35		PRO
36		SER
37		THR
38		TRP
39		TYR
40		VAL
41	DNA	Base-pair
42		Phosphate

Table D.1: Chemically-specific model mapping from LAMMPS atom type ID to the represented particle type.

Interaction	Parameters	Values
E_{RBP}	$\mathbf{K}$	6x6 stiffness matrix for each of the 16 base-pair steps, see table D.6.
	$\boldsymbol{\phi}_0$	6d vector of equilibrium helical parameters for each of the 16 base-pair steps, see table D.6.
E_{Bonds}	r_0	3.5 Å for histone tails, encoded from reference structure for globular domain GNM.
	k	$10\,\text{kcal/mol/Å}^2$
E_{KH}	ϵ, σ, λ	Protein–protein: tables 3.2, D.4, and D.5. DNA–protein: table 3.3. DNA-DNA: no interaction ($\epsilon = 0$).
$E_{\text{Electrostatic}}$	Charge	Protein: table 3.2. DNA: table 3.1.
	κ	Salt-dependence via equation 3.12.

Table D.2: Summary of chemically-specific model parameters.

Interaction	Parameters	Values
$E_{\text{Minimal-RBP}}$	$\mathbf{K}$	diag(0.301,0.235,1.56, 0.00614, 0.00515,0.00724).
	$\boldsymbol{\phi}_0$	$(0, 0, 16.44, 0, 0, 166.69)$.
$E_{\text{Minimal-LJ}}$	ϵ, σ, r_c^{LJ}	Table 5.2.
$E_{\text{Ansiotropic}}$	$\mathbf{S}$, ϵ, r_c, $\mathbf{E}$	Table 5.4.

Table D.3: Summary of minimal model parameters.

Table D.4: KH parameter set A, from Ref. [133].

Amino-acid pair	ϵ	λ
CYS–CYS	0.29889	1
CYS–MET	0.256461	1
MET–MET	0.300776	1
CYS–PHE	0.332833	1
MET–PHE	0.404491	1
PHE–PHE	0.470492	1
CYS–ILE	0.304547	1
MET–ILE	0.353576	1
PHE–ILE	0.430892	1
ILE–ILE	0.402605	1
CYS–LEU	0.335662	1
MET–LEU	0.390348	1
PHE–LEU	0.472378	1
ILE–LEU	0.449749	1
LEU–LEU	0.480864	1
CYS–VAL	0.253632	1
MET–VAL	0.287575	1
PHE–VAL	0.379034	1
ILE–VAL	0.356405	1
LEU–VAL	0.396948	1
VAL–VAL	0.306433	1
CYS–TRP	0.252689	1
MET–TRP	0.309261	1
PHE–TRP	0.366776	1
ILE–TRP	0.330947	1
LEU–TRP	0.364891	1
VAL–TRP	0.274375	1
TRP–TRP	0.263061	1
CYS–TYR	0.178202	1
MET–TYR	0.248918	1
PHE–TYR	0.319633	1
ILE–TYR	0.280975	1
LEU–TYR	0.320576	1
VAL–TYR	0.221574	1
TRP–TYR	0.225346	1
TYR–TYR	0.179145	1
CYS–ALA	0.122573	1
MET–ALA	0.157459	1
PHE–ALA	0.239489	1
ILE–ALA	0.217803	1
LEU–ALA	0.248918	1
VAL–ALA	0.166888	1
TRP–ALA	0.146145	1
TYR–ALA	0.102773	1
ALA–ALA	0.042429	1
CYS–GLY	0.083915	1
MET–GLY	0.105601	1

Continued on next page

Table D.4 – continued from previous page

Amino-acid pair	ϵ	λ
PHE–GLY	0.175374	1
ILE–GLY	0.142373	1
LEU–GLY	0.178202	1
VAL–GLY	0.104659	1
TRP–GLY	0.10843	1
TYR–GLY	0.069772	1
ALA–GLY	0.003771	1
GLY–GLY	0.002829	-1
CYS–THR	0.079201	1
MET–THR	0.116916	1
PHE–THR	0.189517	1
ILE–THR	0.165945	1
LEU–THR	0.195174	1
VAL–THR	0.112202	1
TRP–THR	0.089573	1
TYR–THR	0.069772	1
ALA–THR	0.004714	1
GLY–THR	0.017915	-1
THR–THR	0.014143	-1
CYS–SER	0.055629	1
MET–SER	0.071658	1
PHE–SER	0.165002	1
ILE–SER	0.117859	1
LEU–SER	0.155574	1
VAL–SER	0.073544	1
TRP–SER	0.067887	1
TYR–SER	0.048086	1
ALA–SER	0.024515	-1
GLY–SER	0.042429	-1
THR–SER	0.029229	-1
SER–SER	0.056572	-1
CYS–ASN	0.030172	1
MET–ASN	0.064115	1
PHE–ASN	0.139545	1
ILE–ASN	0.091458	1
LEU–ASN	0.138602	1
VAL–ASN	0.052801	1
TRP–ASN	0.07543	1
TYR–ASN	0.046201	1
ALA–ASN	0.040543	-1
GLY–ASN	0.049972	-1
THR–ASN	0.036772	-1
SER–ASN	0.065058	-1
ASN–ASN	0.055629	-1
CYS–GLN	0.054686	1
MET–GLN	0.097116	1
PHE–GLN	0.172545	1
ILE–GLN	0.132002	1

Continued on next page

Table D.4 – continued from previous page

Amino-acid pair	ϵ	λ
LEU–GLN	0.166888	1
VAL–GLN	0.07543	1
TRP–GLN	0.079201	1
TYR–GLN	0.066001	1
ALA–GLN	0.035829	-1
GLY–GLN	0.057515	-1
THR–GLN	0.034886	-1
SER–GLN	0.073544	-1
ASN–GLN	0.052801	-1
GLN–GLN	0.06883	-1
CYS–ASP	0.0132	1
MET–ASP	0.028286	1
PHE–ASP	0.114087	1
ILE–ASP	0.084858	1
LEU–ASP	0.106544	1
VAL–ASP	0.0198	1
TRP–ASP	0.053744	1
TYR–ASP	0.046201	1
ALA–ASP	0.053744	-1
GLY–ASP	0.064115	-1
THR–ASP	0.044315	-1
SER–ASP	0.060344	-1
ASN–ASP	0.055629	-1
GLN–ASP	0.076372	-1
ASP–ASP	0.099944	-1
CYS–GLU	0.002	-1
MET–GLU	0.058458	1
PHE–GLU	0.12163	1
ILE–GLU	0.094287	1
LEU–GLU	0.124459	1
VAL–GLU	0.037715	1
TRP–GLU	0.067887	1
TYR–GLU	0.049029	1
ALA–GLU	0.071658	-1
GLY–GLU	0.099001	-1
THR–GLU	0.049972	-1
SER–GLU	0.074487	-1
ASN–GLU	0.071658	-1
GLN–GLU	0.080144	-1
ASP–GLU	0.117859	-1
GLU–GLU	0.12823	-1
CYS–HIS	0.125402	1
MET–HIS	0.161231	1
PHE–HIS	0.235717	1
ILE–HIS	0.176317	1
LEU–HIS	0.214031	1
VAL–HIS	0.123516	1
TRP–HIS	0.161231	1

Continued on next page

Table D.4 – continued from previous page

Amino-acid pair	ϵ	λ
TYR–HIS	0.117859	1
ALA–HIS	0.0132	1
GLY–HIS	0.011314	-1
THR–HIS	0.014143	1
SER–HIS	0.015086	-1
ASN–HIS	0.017915	-1
GLN–HIS	0.027343	-1
ASP–HIS	0.004714	1
GLU–HIS	0.011314	-1
HIS–HIS	0.073544	1
CYS–ARG	0.028286	1
MET–ARG	0.080144	1
PHE–ARG	0.161231	1
ILE–ARG	0.12823	1
LEU–ARG	0.165945	1
VAL–ARG	0.07543	1
TRP–ARG	0.107487	1
TYR–ARG	0.083915	1
ALA–ARG	0.041486	-1
GLY–ARG	0.051858	-1
THR–ARG	0.034886	-1
SER–ARG	0.061287	-1
ASN–ARG	0.059401	-1
GLN–ARG	0.044315	-1
ASP–ARG	0.001886	1
GLU–ARG	0.002	-1
HIS–ARG	0.010372	-1
ARG–ARG	0.067887	-1
CYS–LYS	0.030172	-1
MET–LYS	0.0198	1
PHE–LYS	0.102773	1
ILE–LYS	0.069772	1
LEU–LYS	0.103716	1
VAL–LYS	0.020743	1
TRP–LYS	0.039601	1
TYR–LYS	0.031115	1
ALA–LYS	0.090516	-1
GLY–LYS	0.105601	-1
THR–LYS	0.090516	-1
SER–LYS	0.11503	-1
ASN–LYS	0.099944	-1
GLN–LYS	0.092401	-1
ASP–LYS	0.055629	-1
GLU–LYS	0.044315	-1
HIS–LYS	0.086744	-1
ARG–LYS	0.158402	-1
LYS–LYS	0.202717	-1
CYS–PRO	0.07543	1

Continued on next page

Table D.4 – continued from previous page

Amino-acid pair	ϵ	λ
MET–PRO	0.111259	1
PHE–PRO	0.186688	1
ILE–PRO	0.140488	1
LEU–PRO	0.181974	1
VAL–PRO	0.099001	1
TRP–PRO	0.137659	1
TYR–PRO	0.086744	1
ALA–PRO	0.022629	-1
GLY–PRO	0.037715	-1
THR–PRO	0.034886	-1
SER–PRO	0.066001	-1
ASN–PRO	0.069772	-1
GLN–PRO	0.050915	-1
ASP–PRO	0.08863	-1
GLU–PRO	0.09523	-1
HIS–PRO	0.001886	-1
ARG–PRO	0.053744	-1
LYS–PRO	0.122573	-1
PRO–PRO	0.049029	-1

Table D.5: KH parameter set D, from Ref. [133].

CYS–CYS	0.509719	1
CYS–MET	0.448877	1
MET–MET	0.512423	1
CYS–PHE	0.558393	1
MET–PHE	0.661148	1
PHE–PHE	0.75579	1
CYS–ILE	0.517831	1
MET–ILE	0.588137	1
PHE–ILE	0.699005	1
ILE–ILE	0.658443	1
CYS–LEU	0.562449	1
MET–LEU	0.640867	1
PHE–LEU	0.758494	1
ILE–LEU	0.726045	1
LEU–LEU	0.770663	1
CYS–VAL	0.444821	1
MET–VAL	0.493495	1
PHE–VAL	0.624642	1
ILE–VAL	0.592194	1
LEU–VAL	0.650331	1
VAL–VAL	0.520535	1
CYS–TRP	0.443469	1
MET–TRP	0.524592	1
PHE–TRP	0.607066	1
ILE–TRP	0.555688	1
LEU–TRP	0.604362	1
VAL–TRP	0.474566	1
TRP–TRP	0.458342	1
CYS–TYR	0.336658	1
MET–TYR	0.438061	1
PHE–TYR	0.539464	1
ILE–TYR	0.48403	1
LEU–TYR	0.540816	1
VAL–TYR	0.398852	1
TRP–TYR	0.40426	1
TYR–TYR	0.33801	1
CYS–ALA	0.256888	1
MET–ALA	0.306913	1
PHE–ALA	0.424541	1
ILE–ALA	0.393444	1
LEU–ALA	0.438061	1
VAL–ALA	0.320433	1
TRP–ALA	0.290689	1
TYR–ALA	0.228495	1
ALA–ALA	0.141964	1
CYS–GLY	0.201454	1
MET–GLY	0.232551	1
PHE–GLY	0.332602	1

Continued on next page

Table D.5 – continued from previous page

Amino-acid pair	ϵ	λ
ILE–GLY	0.28528	1
LEU–GLY	0.336658	1
VAL–GLY	0.231199	1
TRP–GLY	0.236607	1
TYR–GLY	0.181173	1
ALA–GLY	0.086531	1
GLY–GLY	0.077066	1
CYS–THR	0.194694	1
MET–THR	0.248775	1
PHE–THR	0.352882	1
ILE–THR	0.319081	1
LEU–THR	0.360995	1
VAL–THR	0.242015	1
TRP–THR	0.209566	1
TYR–THR	0.181173	1
ALA–THR	0.087883	1
GLY–THR	0.055434	1
THR–THR	0.060842	1
CYS–SER	0.160893	1
MET–SER	0.183877	1
PHE–SER	0.317729	1
ILE–SER	0.250127	1
LEU–SER	0.304209	1
VAL–SER	0.186582	1
TRP–SER	0.178469	1
TYR–SER	0.150076	1
ALA–SER	0.045969	1
GLY–SER	0.020281	1
THR–SER	0.039209	1
SER–SER	0.002	-1
CYS–ASN	0.124388	1
MET–ASN	0.173061	1
PHE–ASN	0.281224	1
ILE–ASN	0.21227	1
LEU–ASN	0.279872	1
VAL–ASN	0.156837	1
TRP–ASN	0.189286	1
TYR–ASN	0.147372	1
ALA–ASN	0.022985	1
GLY–ASN	0.009464	1
THR–ASN	0.028393	1
SER–ASN	0.012168	-1
ASN–ASN	0.001352	1
CYS–GLN	0.159541	1
MET–GLN	0.220383	1
PHE–GLN	0.328546	1
ILE–GLN	0.270408	1
LEU–GLN	0.320433	1

Continued on next page

Table D.5 – continued from previous page

Amino-acid pair	ϵ	λ
VAL–GLN	0.189286	1
TRP–GLN	0.194694	1
TYR–GLN	0.175765	1
ALA–GLN	0.029745	1
GLY–GLN	0.001352	-1
THR–GLN	0.031097	1
SER–GLN	0.024337	-1
ASN–GLN	0.005408	1
GLN–GLN	0.017577	-1
CYS–ASP	0.100051	1
MET–ASP	0.121684	1
PHE–ASP	0.244719	1
ILE–ASP	0.202806	1
LEU–ASP	0.233903	1
VAL–ASP	0.109515	1
TRP–ASP	0.158189	1
TYR–ASP	0.147372	1
ALA–ASP	0.004056	1
GLY–ASP	0.010816	-1
THR–ASP	0.017577	1
SER–ASP	0.005408	-1
ASN–ASP	0.001352	1
GLN–ASP	0.028393	-1
ASP–ASP	0.062194	-1
CYS–GLU	0.081122	1
MET–GLU	0.164949	1
PHE–GLU	0.255536	1
ILE–GLU	0.216326	1
LEU–GLU	0.259592	1
VAL–GLU	0.135204	1
TRP–GLU	0.178469	1
TYR–GLU	0.151428	1
ALA–GLU	0.021633	-1
GLY–GLU	0.060842	-1
THR–GLU	0.009464	1
SER–GLU	0.025689	-1
ASN–GLU	0.021633	-1
GLN–GLU	0.033801	-1
ASP–GLU	0.087883	-1
GLU–GLU	0.102755	-1
CYS–HIS	0.260944	1
MET–HIS	0.312321	1
PHE–HIS	0.419132	1
ILE–HIS	0.333954	1
LEU–HIS	0.388035	1
VAL–HIS	0.25824	1
TRP–HIS	0.312321	1
TYR–HIS	0.250127	1

Continued on next page

Table D.5 – continued from previous page

Amino-acid pair	ϵ	λ
ALA–HIS	0.100051	1
GLY–HIS	0.064898	1
THR–HIS	0.101403	1
SER–HIS	0.05949	1
ASN–HIS	0.055434	1
GLN–HIS	0.041913	1
ASP–HIS	0.087883	1
GLU–HIS	0.064898	1
HIS–HIS	0.186582	1
CYS–ARG	0.121684	1
MET–ARG	0.196046	1
PHE–ARG	0.312321	1
ILE–ARG	0.265	1
LEU–ARG	0.319081	1
VAL–ARG	0.189286	1
TRP–ARG	0.235255	1
TYR–ARG	0.201454	1
ALA–ARG	0.021633	1
GLY–ARG	0.00676	1
THR–ARG	0.031097	1
SER–ARG	0.00676	-1
ASN–ARG	0.004056	-1
GLN–ARG	0.017577	1
ASP–ARG	0.083826	1
GLU–ARG	0.081122	1
HIS–ARG	0.06625	1
ARG–ARG	0.016224	-1
CYS–LYS	0.037857	1
MET–LYS	0.109515	1
PHE–LYS	0.228495	1
ILE–LYS	0.181173	1
LEU–LYS	0.229847	1
VAL–LYS	0.110867	1
TRP–LYS	0.137908	1
TYR–LYS	0.12574	1
ALA–LYS	0.048673	-1
GLY–LYS	0.070306	-1
THR–LYS	0.048673	-1
SER–LYS	0.083826	-1
ASN–LYS	0.062194	-1
GLN–LYS	0.051378	-1
ASP–LYS	0.001352	1
GLU–LYS	0.017577	1
HIS–LYS	0.043265	-1
ARG–LYS	0.14602	-1
LYS–LYS	0.209566	-1
CYS–PRO	0.189286	1
MET–PRO	0.240663	1

Continued on next page

Table D.5 – continued from previous page

Amino-acid pair	ϵ	λ
PHE–PRO	0.348826	1
ILE–PRO	0.282576	1
LEU–PRO	0.342066	1
VAL–PRO	0.223087	1
TRP–PRO	0.27852	1
TYR–PRO	0.20551	1
ALA–PRO	0.048673	1
GLY–PRO	0.027041	1
THR–PRO	0.031097	1
SER–PRO	0.01352	-1
ASN–PRO	0.018929	-1
GLN–PRO	0.008112	1
ASP–PRO	0.045969	-1
GLU–PRO	0.055434	-1
HIS–PRO	0.078418	1
ARG–PRO	0.004056	1
LYS–PRO	0.094643	-1
PRO–PRO	0.010816	1

Table D.6: Helical parameters for the DNA rigid base-pair potential, from Ref [76].

Base-pair step	Shift	Slide	Rise	Tilt	Roll	Twist
AA-TT						
ϕ_0	-0.3	-0.3	3.3	-2.6	0.3	35.4
K	1.72017	0.19796	0.32533	-0.01249	0.00576	0.05913
	0.19797	2.12618	0.75074	-0.00581	-0.05309	-0.10162
	0.32534	0.75074	7.64359	-0.18348	-0.04547	-0.1485
	-0.01249	-0.00581	-0.18349	0.03738	0.00211	0.00597
	0.00576	-0.05309	-0.04547	0.00211	0.01961	0.00742
	0.05913	-0.10162	-0.1485	0.00597	0.00742	0.02761
AC-GT						
ϕ_0	0.1	-0.6	3.3	-0.7	-0.6	32
K	1.28288	0.13127	0.29502	-0.0278	0.00302	0.03646
	0.13127	2.97699	2.10518	-0.0228	0.03038	-0.10881
	0.29502	2.10518	8.83137	0.04907	0.10478	-0.14741
	-0.0278	-0.0228	0.04907	0.03776	0.00378	0.00418
	0.00302	0.03038	0.10479	0.00378	0.02306	0.00708
	0.03646	-0.10881	-0.14742	0.00418	0.00708	0.03551
AG-CT						
ϕ_0	-0.4	-0.6	3.4	-2.5	3.1	33.5
K	1.3999	0.27887	0.27572	-0.03917	0.0208	0.07408
	0.27887	1.78493	0.99427	-0.00395	-0.01181	-0.06894
	0.27572	0.99427	7.0413	-0.15259	-0.02113	-0.14069
	-0.03917	-0.00395	-0.15259	0.03699	0.0041	0.0063

Continued on next page

Table D.6 – continued from previous page

Base-pair step	Shift	Slide	Rise	Tilt	Roll	Twist
	0.0208	-0.01181	-0.02113	0.0041	0.01912	0.005
	0.07408	-0.06894	-0.14069	0.0063	0.005	0.02797
AT-AT						
ϕ_0	0	-0.8	3.3	0	-0.5	30.4
K	1.04661	0	0	0.03066	0	0
	0	3.76524	2.16877	0	-0.03029	-0.04352
	0	2.16877	9.33735	0	0.06011	-0.09916
	0.03066	0	0	0.03494	0	0
	0	-0.03029	0.06011	0	0.02157	0.00877
	0	-0.04352	-0.09916	0	0.00877	0.03122
CA-TG						
ϕ_0	-0.2	-0.2	3.1	0.2	10.3	29.6
K	1.05328	0.07552	0.23196	-0.03375	0.00158	-0.01277
	0.07552	1.79677	0.52999	-0.00472	0.02114	-0.03759
	0.23197	0.53	6.30474	0.01221	-0.07708	-0.144
	-0.03375	-0.00472	0.01221	0.02498	-0.00127	0.00137
	0.00158	0.02114	-0.07708	-0.00127	0.01647	0.00541
	-0.01277	-0.03759	-0.14399	0.00137	0.00541	0.01534
CC-GG						
ϕ_0	0.2	-0.7	3.5	0	4.9	32.7
K	1.43205	-0.29768	-0.35747	-0.08468	-0.01448	-0.03982
	-0.29768	1.5731	1.18241	-0.00677	0.01068	-0.09303
	-0.35748	1.18243	7.85985	0.25771	0.00156	-0.11971
	-0.08468	-0.00677	0.25771	0.04202	0.00134	0.00022
	-0.01448	0.01068	0.00156	0.00134	0.02012	0.00493
	-0.03982	-0.09303	-0.1197	0.00022	0.00493	0.02602
CG-CG						
ϕ_0	0	0	3.1	0	9.3	29.8
K	1.05459	0	0	-0.07841	0	0
	0	1.91048	0.56341	0	0.02388	-0.05037
	0	0.56342	6.11191	0	-0.04625	-0.14832
	-0.07841	0	0	0.02587	0	0
	0	0.02388	-0.04625	0	0.01562	0.00319
	0	-0.05037	-0.14832	0	0.00319	0.014
CT-AG						
ϕ_0	0.4	-0.6	3.4	2.5	3.1	33.5
K	1.3999	-0.27887	-0.27572	-0.03917	-0.0208	-0.07408
	-0.27887	1.78493	0.99427	0.00395	-0.01181	-0.06894
	-0.27572	0.99427	7.0413	0.15259	-0.02113	-0.14069
	-0.03917	0.00395	0.15259	0.03699	-0.0041	-0.0063
	-0.0208	-0.01181	-0.02113	-0.0041	0.01912	0.005

Continued on next page

Table D.6 – continued from previous page

Base-pair step	Shift	Slide	Rise	Tilt	Roll	Twist
	-0.07408	-0.06894	-0.14069	-0.0063	0.005	0.02797
GA-TC						
ϕ_0	-0.4	-0.3	3.4	-1.6	1.6	36.6
K	1.31766	0.29517	0.40655	-0.04665	0.00977	0.01782
	0.29517	1.87898	1.00952	0.00116	-0.01328	-0.09736
	0.40656	1.00952	8.48201	-0.25663	-0.01013	-0.12376
	-0.04665	0.00116	-0.25663	0.03758	-0.00252	0.00117
	0.00977	-0.01328	-0.01012	-0.00252	0.02025	0.00861
	0.01782	-0.09736	-0.12376	0.00117	0.00861	0.02441
GC-GC						
ϕ_0	0	-0.4	3.5	0	-1.3	35.7
K	1.179	0	0	-0.08357	0	0
	0	2.58821	2.05067	0	0.08753	-0.07364
	0	2.05063	9.46559	0	0.1572	-0.18431
	-0.08357	0	0	0.03618	0	0
	0	0.08753	0.1572	0	0.02569	0.00442
	0	-0.07364	-0.18431	0	0.00442	0.02238
GG-CC						
ϕ_0	-0.2	-0.7	3.5	0	4.9	32.7
K	1.43205	0.29768	0.35747	-0.08468	0.01448	0.03982
	0.29768	1.5731	1.18241	0.00677	0.01068	-0.09303
	0.35748	1.18243	7.85985	-0.25771	0.00156	-0.11971
	-0.08468	0.00677	-0.25771	0.04202	-0.00134	-0.00022
	0.01448	0.01068	0.00156	-0.00134	0.02012	0.00493
	0.03982	-0.09303	-0.1197	-0.00022	0.00493	0.02602
GT-AC						
ϕ_0	-0.1	-0.6	3.3	0.7	-0.6	32
K	1.28288	-0.13127	-0.29502	-0.0278	-0.00302	-0.03646
	-0.13127	2.97699	2.10518	0.0228	0.03038	-0.10881
	-0.29502	2.10518	8.83137	-0.04907	0.10478	-0.14741
	-0.0278	0.0228	-0.04907	0.03776	-0.00378	-0.00418
	-0.00302	0.03038	0.10479	-0.00378	0.02306	0.00708
	-0.03646	-0.10881	-0.14742	-0.00418	0.00708	0.03551
TA-TA						
ϕ_0	0	-0.2	3.2	0	10	28.9
K	0.64315	0	0	-0.01544	0	0
	0	1.24864	0.53653	0	0.02038	-0.04224
	0	0.53653	6.07971	0	-0.07916	-0.14834
	-0.01544	0	0	0.019	0	0
	0	0.02038	-0.07916	0	0.01464	0.00831
	0	-0.04224	-0.14835	0	0.00831	0.01759

Continued on next page

Table D.6 – continued from previous page

Base-pair step	Shift	Slide	Rise	Tilt	Roll	Twist
TC-GA						
ϕ_0	0.4	-0.3	3.4	1.6	1.6	36.6
K	1.31766	-0.29517	-0.40655	-0.04665	-0.00977	-0.01782
	-0.29517	1.87898	1.00952	-0.00116	-0.01328	-0.09736
	-0.40656	1.00952	8.48201	0.25663	-0.01013	-0.12376
	-0.04665	-0.00116	0.25663	0.03758	0.00252	-0.00117
	-0.00977	-0.01328	-0.01012	0.00252	0.02025	0.00861
	-0.01782	-0.09736	-0.12376	-0.00117	0.00861	0.02441
TG-CA						
ϕ_0	0.2	-0.2	3.1	-0.2	10.3	29.6
K	1.05328	-0.07552	-0.23196	-0.03375	-0.00158	0.01277
	-0.07552	1.79677	0.52999	0.00472	0.02114	-0.03759
	-0.23197	0.53	6.30474	-0.01221	-0.07708	-0.144
	-0.03375	0.00472	-0.01221	0.02498	0.00127	-0.00137
	-0.00158	0.02114	-0.07708	0.00127	0.01647	0.00541
	0.01277	-0.03759	-0.14399	-0.00137	0.00541	0.01534
TT-AA						
ϕ_0	0.3	-0.3	3.3	2.6	0.3	35.4
K	1.72017	-0.19796	-0.32533	-0.01249	-0.00576	-0.05913
	-0.19797	2.12618	0.75074	0.00581	-0.05309	-0.10162
	-0.32534	0.75074	7.64359	0.18348	-0.04547	-0.1485
	-0.01249	0.00581	0.18349	0.03738	-0.00211	-0.00597
	-0.00576	-0.05309	-0.04547	-0.00211	0.01961	0.00742
	-0.05913	-0.10162	-0.1485	-0.00597	0.00742	0.02761